Überprüfung und Anwendung von Multilevel-Messmodellen für Fragebögen zur Lehrveranstaltungsevaluation

Dissertation

zur Erlangung des akademischen Grades

doctor philosophiae (Dr. phil.)

vorgelegt dem Rat der Fakultät für Sozial- und Verhaltenswissenschaften

der Friedrich-Schiller-Universität Jena

von Dipl.-Psych. Erik Sengewald

geboren am 18. März 1985 in Plauen

Gutachter

1. Prof. Dr. Rolf Steyer, Jena

2. Prof. Dr. Matthias Ziegler, Berlin

Tag der mündlichen Prüfung: 27. Januar 2016

Für Jan, Julian, Marie-Ann

und unsere gemeinsame Zukunft

Danksagung

In den vergangenen sechs Jahren begleiteten mich viele Menschen auf meinem beruflichen und privaten Weg, denen ich an dieser Stelle danken möchte. Besonderer Dank gilt Prof. Dr. Steyer. Er hat als Betreuer dieser Arbeit nicht nur hilfreiche fachliche Unterstützung geleistet, sondern vor allem den geschützten Rahmen hergestellt, in dem es sich hervorragend arbeiten ließ. Prof. Dr. Steyer begleitete mich seit Beginn meines Studium vor 11 Jahren und ermöglichte diese Arbeit auch durch seinen Enthusiasmus gegenüber der Statistik. Dieser weckte auch in mir das Interesse und katalysierte meinen beruflichen Werdegang in eine Richtung, mit der ich sehr zufrieden bin. Vielen Dank dafür. Weiterhin danke ich Prof. Dr. Ziegler. Er zeigte mir eine neue Perspektive für meine Dissertation auf, als er mich nach einem Vortrag in seiner Arbeitsgruppe in das Themenheft „Lehrevaluationen“ in die Zeitschrift Diagnostica einlud. Der damit geschaffene Anreiz beschleunigte die Fertigstellung der Dissertation. Ganz herzlichen Dank.

Während meiner beruflichen Laufbahn und auch als Student traf ich auf wichtige Menschen, die mich durch ihre vertrauensvolle Haltung mir gegenüber stets forderten und förderten. Nur durch derartige Unterstützung kann ein Ziel, wie die Promotion, über lange Zeit bestehen. Besonders dankbar bin ich dafür Hendryk Böhme. Bereits während meines Studiums unterstützte er mich an entscheidenden Stellen und prägte damit meinen beruflichen Werdegang wie kein anderer. Ihm danke ich für die positiven Erfahrungen bei der Arbeit in seinem Projekt, die Unterstützung bei der Entscheidung zur Daimler AG zu gehen und dort meine Diplomarbeit zu schreiben und für weitere Momente, in denen er mir Chancen und Möglichkeiten aufzeigte, die meinen Lebensweg prägen sollten. Die vergangenen sechs Jahre, die ich an meiner Dissertation arbeitete, waren vor allem durch viel Projektarbeit geprägt, sodass die Zeit für diese wissenschaftliche Arbeit manchmal zu kurz kam. Dennoch gelang es, die Dissertation nicht aus den Augen zu verlieren und in der Wissenschaft Fuß zu fassen. Hierfür danke ich vor allem Anja Vetterlein. Als Büroehefrau und als Wissenschaftlerin verstand ich sie immer als Partnerin auf einem gemeinsamen Weg. Gemeinsam nahmen wir wissenschaftliche Hürden und standen für unsere Interessen ein. Ihre moralische und fachliche Unterstützung weiß ich bis heute zu schätzen. Besonders ihre gewissenhafte Haltung gegenüber der Formatierung von Tabellen und ihre Abneigung gegenüber sprachlicher Vielfalt in Form verschiedener Schriftarten gaben mir einen Feinschliff in Sachen Arbeitsweise, der für meine tägliche Arbeit heute sehr nützlich ist. Für ih-

re selbstlose Unterstützung in der finalen Phase dieser Arbeit gilt mein Dank Juliane Brachwitz. Sie unterstützte mich als es für mich am wichtigsten war und half mir damit auch am Ende dieses Weges nicht zu verzweifeln. Ganz herzlichen Dank dafür. Besonderer Dank gilt auch meinen Kollegen aus verschiedenen beruflichen Stationen. Stefanie Dubiella, Hans-Josef Küting und Florian Hanisch begeisterten mich bei der Daimler AG für die angewandte psychologische Forschung und Entwicklung. Katrin Schaller, Marcel Bauer, Norman Rose, Ulf Kröhne, Christiane Fiege, Jan Marten Ihme, Sonja Hahn, Steffi Pohl und Sven Hartenstein prägten meine 10 Jahre am Lehrstuhl für Methodenlehre und unterstützten mich jederzeit.

Das Vertrauen meiner Familie ermöglichte es mir, diese Arbeit intrinsisch motiviert und ohne äußeren Leistungsdruck anzufertigen. Dafür möchte ich mich ganz herzlich bei meinen Eltern Elke und Ralf Sengewald und bei meinem Bruder Kay Sengewald bedanken. Mit stets offenen Armen und bedingungsloser Akzeptanz meiner Entscheidungen, durfte ich selbst herausfinden was ich möchte und welchen Weg ich einschlage. Warme Unterstützung und ein glückliches Umfeld waren mir immer sicher, egal wohin mich der Weg führte. Danke! Der Weg führte mich bald in die Arme von Marie-Ann. Ihr verdanke ich nicht nur zwei Kinder und viele schöne Stunden und Tage zusammen, ihr verdanke ich vor allem ein Leben in Liebe, Harmonie, manchmal emotional fachlichen Auseinandersetzungen, mit viel Abwechslung und Freuden und immer Glück. In den entscheidenden Phasen meiner Dissertation war sie zudem mehr als eine moralische Stütze. Ihr verdanke ich wesentliche Änderungsvorschläge, viele fachliche Diskussionen und Hinweise, die allesamt, wenn auch manchmal mit Widerstand, in die Arbeit eingeflossen sind. Sie entschied sich früh für eine gemeinsame Familie und nahm mir damit die Möglichkeit auch am Wochenende an der Dissertation zu arbeiten. Vielen Dank dafür! Unsere Söhne zeigen mir stets aufs Neue, was wichtig ist im Leben. Jan und Julian können mir zwar keinen Doktortitel verleihen, aber für sie bin ich schon immer Doktor, Feuerwehrmann, Polizist, Bauarbeiter, Löwenbändiger, Schaukelpferd, Kuscheltier und vieles mehr, für das man vor allem Zuneigung und Zeit benötigt. Danke Marie-Ann, für dieses Geschenk und dafür, dass alles möglich ist, was wir vier erreichen wollen.

Abstract

Students' evaluation of teaching (SET) is a common method for feedback at universities. According to Marsh (2007b) SETs are collected to provide diagnostic feedback to teachers for improving teaching. However, SETs are criticised for not meeting psychometric criteria of confirmatory factor analysis (CFA) appropriately (Marsh et al., 2009). The present dissertation addresses several topics of applying CFA to SET questionnaires. In the first two sections of this work, I introduce the challenges when CFA is applied and show how SETs are implemented at the university of Jena. This is followed by a study about the appropriate method of CFA. The study compares two common types of CFA (using either student ratings or class means) with the multilevel confirmatory factor analysis (ML-CFA) based on an empirical example of 183 334 student ratings. Due to the results the application of a ML-CFA to SETs is strongly recommended. Another challenge when studying the structural model of SET questionnaires are multiple evaluations. It is possible that students evaluate more than one course whithin a semester or during their studies. Thus, they appear in several SET results of different courses. The results of this study shows that multiple evaluations impair the fit of the model depending of the type of multiple evaluation and the CFA method used. The ML-CFA, which appears to meet the data best compared to the conventional methods, is used to examine the effect of report quality on the course quality. The third study shows the effect of two reports on the course quality in a randomized experiment with $N = 283$ lecturers. Course quality is measured by the questionnaire PELVE (Process and Outcome Oriented Students' Evaluation of Teaching, Born, Loßnitzer & Schmidt, 2006) on seven latent dimensions using the above mentioned multilevel measurement model. Despite the complexity of the new report, the study shows a positive effect on the dimension *course material*. This work finishes with a section discussing the results of the studies, stating that SETs are a useful tool for feedback of the course quality to the lecturer. However, it is especially important to decide which method is appropriate and how the structure of the sample can be considered when evaluating the quality of a SET questionnaire.

Zusammenfassung

Die Lehrveranstaltungsevaluation (LVE) ist ein probates und oft eingesetztes Mittel zur Erfassung der Qualität einer Lehrveranstaltung. In erster Linie werden Fragebögen zur LVE als Feedbackinstrument für den Dozenten verwendet, mit dem Ziel, die Lehre zu verbessern Marsh (2007b). Allerdings wird die psychometrische Qualität der eingesetzten Fragebögen oft angezweifelt (vgl. Marsh et al., 2009). Die vorliegende Arbeit beschäftigt sich mit verschiedenen Herausforderungen, die in Zusammenhang mit der Überprüfung der psychometrischen Qualität stehen. In den ersten beiden Kapiteln dieser Arbeit wird darauf genauer eingegangen und die LVE an der Friedrich-Schiller-Universität Jena vorgestellt. Die anschließende Studie thematisiert die Eignung verschiedener konventioneller konfirmatorischer Faktorenanalysen (CFA) zur Überprüfung der Modellgüte und vergleicht sie mit der Multi-Level-Faktorenanalyse (ML-CFA) unter Verwendung einer Stichprobe mit 183 334 Studentenurteilen. Aus dem Vergleich dieser CFA-Verfahren lässt sich eine klare Empfehlung für die Verwendung der ML-CFA ableiten. Eine weitere Herausforderung bei der Überprüfung der psychometrischen Qualität von LVE-Fragebögen sind Mehrfachevaluationen. Sie entstehen, wenn Studenten mehrere Veranstaltungen evaluieren und damit mehrfach in den Analysestichproben vorkommen. Die Ergebnisse der zweiten Studie verdeutlichen, dass Mehrfachevaluationen einen Einfluss auf die Ergebnisse von CFA haben und dass sich dieser zwischen den verschiedenen CFA-Verfahren unterscheidet. Die ML-CFA, welche nach den Ergebnissen beider Studien am besten zur Überprüfung des Messmodells geeignet ist, wird in der dritten Studie zur Evaluation unterschiedlicher Berichte eingesetzt. Diese Studie zeigt mit Hilfe eines Experiments mit $N = 283$ Dozenten den Effekt verschiedener Darstellungsformen im Ergebnisbericht auf die Qualität der Lehrveranstaltung. Die Lehrveranstaltungsqualität wird hier mit dem Fragebogen PELVE (Prozess- und Ergebnisorientierte Lehrveranstaltungsevaluation, Born et al., 2006) auf sieben Dimensionen durch das bereits erwähnte Multi-Level-Messmodell erfasst. Im Ergebnis zeigt sich ein positiver Effekt einer komplexeren Darstellungsform gegenüber der anderen auf der Dimension Begleitmaterialien. Die Arbeit schließt mit einer Diskussion der Ergebnisse aller drei Studien ab und verdeutlicht die empirischen Belege für die Nützlichkeit der LVE als Feedbackinstrument. Darüber hinaus wird die Bedeutung der Methode und der Analysestichprobe unterstrichen, die für die Überprüfung der psychometrischen Qualität von LVE-Fragebögen eingesetzt werden.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Wissenschaftliche Forschungsarbeiten zur Evaluation von Lehrveranstaltungen liegen bereits seit 1927 vor (Remmers & Brandenburg, 1927). Aktuell gewinnt die Lehrveranstaltungsevaluation (LVE) zunehmend an Bedeutung. Vor allem im Rahmen der Akkreditierung von Hochschulen stellt sie meist einen wichtigen Baustein des Qualitätssicherungssystems an Hochschulen dar. Für die Verwendung als Instrument zur Qualitätssicherung auf Hochschulebene oder gar als Mittel zur Steuerung, wurden Fragebögen zur LVE jedoch ursprünglich nicht entwickelt. Im Fokus stand vorrangig die Verwendung der LVE-Fragebögen als Feedbackinstrument. Der Lehrende soll damit einen Überblick über die Situation in seiner Veranstaltung erhalten und Stärken sowie Schwächen seiner Lehre identifizieren können. Hierfür wurden bis heute viele verschiedene Fragebögen zur Evaluation der Lehre an Hochschulen entwickelt. An den Hochschulen wurden Fragebögen konstruiert, die genau auf ihre Bedürfnisse abgestimmt sind. Die Fragebögen zur LVE basieren auf unterschiedlichen Theorien zur Lehrveranstaltungsqualität (LVQ) und beanspruchen unterschiedliche Facetten dieser Qualität zu messen. Ziel der Fragebogenentwicklung ist es, für die LVE ein Messinstrument zu entwickeln mit dem reliable Aussagen bezüglich der zu messenden Konstrukte getroffen werden können. Ob mit Hilfe der entwickelten Items eines Fragebogens tatsächlich die intendierten Konstrukte gemessen werden können, ist eine Fragestellung, die über die hier vorliegende Arbeit hinaus geht und im Allgemeinen mit Validitätsstudien beantwortet werden kann.

Gegenstand dieser Arbeit ist die Betrachtung der Methoden zur Überprüfung der psychometrischen Qualität bzw. der Dimensionalität von LVE-Fragebögen. Hierfür sind zwei Studien in die Arbeit eingebunden, die die eher traditionellen konfirmatorischen Faktorenanalysen mit der Multilevel-Faktorenanalyse vergleichen und dabei verschiedene Besonderheiten im Rahmen der LVE berücksichtigen (siehe Manuskript 1 und Manuskript 2 in Abschnitt 4 und Abschnitt 5). Die dritte Studie verdeutlicht im Rahmen einer Treatmentstudie, wie die gewonnen Erkenntnisse zukünftig für die Hochschulforschung zur LVE verwendet werden können. Die drei Studien wurden im Universitätsprojekt Lehrevaluation des Lehrstuhls für Methodenlehre und Evaluationsforschung der Frie-

drich-Schiller-Universität Jena (FSU Jena) durchgeführt [1]. Im folgenden werden die Titel der Studien, die auch in die vorliegende Arbeit eingebunden sind, genannt.

1 Multilevel Faktorenanalyse für Fragebögen zur Lehrveranstaltungsevaluation

2 Einfluss der Mehrfachevaluation auf die Ergebnisse konfirmatorischer Faktorenanalysen in der Lehrveranstaltungsevaluation

3 Ergebnisdarstellung in der Lehrveranstaltungsevaluation: Effekte verschiedener Berichte auf die Qualität von Lehrveranstaltungen.

Für die Feedbackfunktion der LVE sind konfirmatorische Faktorenanalysen zunächst nicht nötig. Hier erhält der Lehrende auf Itemebene einen Evaluationsbericht, der die Urteile der Studenten in Form von Verteilungsparametern zusammenfasst. Für eine weiterführende Verwendung der LVE-Ergebnisse im Rahmen der Hochschulforschung und die Auswertung auf aggregierter Ebene ist Wissen über die faktorielle Struktur des Fragebogens für die Interpretation der Ergebnisse nötig. Konfirmatorische Faktorenanalysen (CFA) für Fragebögen zur Lehrveranstaltungevaluation werden zur Überprüfung des theoretischen Messmodells der Fragebögen eingesetzt. Die CFA-Ergebnisse zeigen oft keinen zufriedenstellenden Modellfit (vgl. Marsh et al., 2009), sodass häufig auf exploratorische Verfahren oder auf datengeleitete Veränderungen des Messmodells zurückgegriffen wird. Als Reaktion auf schlecht passende Modelle werden einzelne Items verändert oder entfernt, es werden Doppelladungen von Items in den Messmodellen zugelassen oder die Multidimensionalität der Fragebögen wird nicht berücksichtigt. Alternativ werden auch exploratorische Faktorenanalysen eingesetzt, die jedoch keine inferenzstatistische Prüfung eines theoretischen Messmodells erlauben.

In der vorliegenden Arbeit werden unterschiedliche CFA-Verfahren und strukturelle Eigenschaften der LVE-Stichproben als Einflussgrößen für die Modellpassung untersucht. Kernstück der drei Studien ist die Verwendung eines im Kontext der LVE bisher nur selten eingesetzten Verfahrens zur CFA in der LVE, die konfirmatorische Multilevel-Analyse (ML-CFA). Studie 1 (vgl. Manuskript 1 in Kapitel 4) vergleicht die ML-CFA mit den konventionellen Methoden der CFA, die entweder nur die Studentenebene oder die Veranstaltungsebene berücksichtigen. Unter dem Begriff „Studentenebene“ versteht man an dieser Stelle die Analyse der Rohdaten, die durch Antworten der Studenten einer Veranstaltung auf Items der LVE zustande kommen. Bei Analysen auf „Veranstaltungsebene“ werden indes Werte der Veranstaltung auf diesen Items verwendet. Hierfür werden die Studentenurteile

[1] Diese Arbeit wurde im Rahmen des gemeinsamen Bund-Länder-Programms für bessere Studienbedingungen und mehr Qualität in der Lehre aus Mitteln des Bundesministeriums für Bildung und Forschung (BMBF) unter dem Förderkennzeichen 01PL12071 gefördert.

innerhalb einer Veranstaltung im Vorfeld der Analyse aggregiert. Im Fokus der vorliegenden Arbeit steht ein empirischer Vergleich der Passung des theoretischen Messmodells eines Fragebogens zu den LVE-Daten für verschiedener CFA-Verfahren. Durch die Verwendung verschiedener CFA-Verfahren unterscheiden sich die konstruierten latenten Variablen, sodass die vom Modell implizierten Varianz-Kovarianz-Matrizen und damit auch die Messmodelle, die geprüft werden, unterschiedlich sind. Den Messmodellen liegt jedoch stets die gleiche Zuordnung der Item zu latenten Variablen zugrunde. Damit basieren die Messmodelle alle auf demselben theoretischen Gerüst des eingesetzten Fragebogens.

Die Studien liefern einen wesentlichen Beitrag zur Diskussion über die Qualität von Fragebögen zur LVE. Bisher wird die Qualität bzw. der Nutzen dieser Fragebögen oft aufgrund schlechter Ergebnisse der Faktorenanalysen angezweifelt (Spooren, Brockx & Mortelmans, 2013). Vergleicht man die CFA-Verfahren, deuten die unterschiedlichen Ergebnisse auf die Notwendigkeit hin, die Verfahren bezüglich ihrer Eignung zur Prüfung des Messmodells zu hinterfragen. In Studie 1 (vgl. Manuskript 1 in Abschnitt 4) erfolgt deshalb sowohl die empirische als auch die theoretische Gegenüberstellung der unterschiedlichen Modelle sowie die Ableitung von Hinweisen für die Anwendung von CFA-Verfahren im Kontext der LVE. Während Studie 1 mit der ML-CFA die Gruppierung von Studenten in Veranstaltungen bei der CFA berücksichtigt, wird in Studie 2 eine weitere Quelle für die unzureichende Modellpassung untersucht. Hierbei handelt es sich um Mehrfachevaluationen. Der Begriff *Mehrfachevaluation* findet in der Literatur zur LVE bisher keine Verwendung. Er soll eine strukturelle Eigenschaft von Stichproben der LVE zum Ausdruck bringen, die der CFA zugrunde liegen. Mehrfachevaluation kann auf Studentenebene vorliegen, wenn ein Student mehr als eine Veranstaltung evaluiert und auf Dozentenebene, wenn mehrere Veranstaltungen eines Dozenten in der Stichprobe enthalten sind. In Abgrenzung zum Begriff der Messwiederholung handelt es sich bei der Mehrfachevaluation nicht um die Wiederholung einer konkreten Messung, sondern um die Evaluation verschiedener Veranstaltungen durch die gleichen Studenten. Die beobachtbaren Konstellationen der Mehrfachevaluation können beliebig komplex sein. Die LVE einer konkreten Veranstaltung kann von Studenten durchgeführt worden sein, die bereits andere Veranstaltungen evaluiert haben, wobei meist nicht alle Studenten mehrere Veranstaltungen evaluiert haben. Konkrete Stichproben, die für eine CFA verwendet werden, beinhalten meist mehrere LVE, sodass die Mehrfachevaluation bei der Überprüfung des Messmodells zu Abhängigkeiten in den Daten führen kann. Sind diese Abhängigkeiten substantiell, können sie die Modellpassung negativ beeinflussen, sofern sie nicht bei der CFA berücksichtigt werden(vgl. Skinner, Holt & Smith, 1989; Wu & Kwok, 2012). Studie 2 unter-

sucht den Einfluss der Mehrfachevaluation auf die Modellpassung im Kontext der LVE-Daten. Ermöglicht wird diese Betrachtung durch die Realisation verschiedener Stichprobentypen mit bzw. ohne Mehrfachevaluation auf Studenten- und Dozentenebene (vgl. Manuskript 2 in Kapitel 5). Dieser Vergleich verschiedener Konstellationen der Mehrfachevaluation wird in Studie 2 für unterschiedliche CFA-Verfahren, die auch in Studie 1 verwendet werden, durchgeführt. Im Ergebnis zeigen sich verschieden große Unterschiede zwischen den Modellpassungen in Abhängigkeit des gewählten CFA-Verfahrens und der Art der Mehrfachevaluation. Bei einer Mehrfachevaluation auf Studentenebene verschlechtert sich der Modellfit im Vergleich zu Stichproben ohne Mehrfachevaluation vor allem bei der CFA auf Veranstaltungsebene. Welche Konsequenzen dies für die Verwendung der CFA auf Veranstaltungsebene hat und inwiefern die ML-CFA besser geeignet ist, wird in Studie 2 diskutiert.

Während Studie 1 und Studie 2 vorrangig methodische Herausforderungen der CFA im Rahmen der LVE in den Vordergrund stellen, fokussiert die dritte Studie eine Anwendung der ML-CFA in einer Treatmentevaluation zur Verbesserung der Lehrveranstaltungsqualität. Gegenstand des Treatments ist die Optimierung der Ergebnisberichte, die nach der Evaluation einer Veranstaltung an den Dozenten versendet werden. Es wird untersucht, inwiefern die Manipulation der Ergebnisberichte einen Einfluss auf die Rezeption der Ergebnisse hat. Im Detail wird betrachtet, ob sich ein Effekt auf die Lehrveranstaltungsqualität für die nachfolgende Veranstaltung des Dozenten zeigt. Weil die Lehrveranstaltungen in verschiedenen Semestern unterschiedliche Studenten beinhalten, ist eine Methode zur Schätzung der Veranstaltungsqualität nötig, die weitgehend unabhängig von den konkret evaluierenden Studenten ist. Mit der ML-CFA steht ein Verfahren zur Verfügung, das eine Schätzung der Qualität von Veranstaltungen erlaubt, die zu unterschiedlichen Zeitpunkten stattfinden und durch verschiedene Studenten evaluiert werden. Der Vergleich zwischen Kontroll- und Treatmentgruppe erfolgt in Studie 3 auf Basis der geschätzten Faktorwerte der latenten Variablen auf Veranstaltungsebene. Um konfundierende Einflüsse weiterer Variablen zu minimieren und auf theoretischer Ebene auszuschließen, wurde in Studie 3 ein Experiment durchgeführt und eine randomisierte Zuweisung der Dozenten in eine Kontrollgruppe mit einem konventionellen Ergebnisbericht und eine Treatmentgruppe mit einem komprimierten Ergebnisbericht durchgeführt. Beide Varianten des Ergebnisberichts beinhalten die gleiche Information, stellen diese aber unterschiedlich grafisch dar. Ob diese vergleichsweise geringe Intervention Effekte auf die Evaluationsergebnisse der nachfolgenden Veranstaltung des Dozenten hat, wird in Studie 3 untersucht (vgl. Manuskript 3 in Kapitel 6).

Im folgenden Kapitel wird das Konzept der Lehrveranstaltungsevaluation genauer eingeführt und der aktuelle Forschungsstand kurz dargestellt. Anschließend werden der Evaluationsablauf und

die Fragebogenentwicklung am Besipiel der FSU Jena genauer erläutert. Die genannten Studien sind anschließend eingebunden. Ein abschließendes Kapitel fasst die Ergebnisse der Studien zusammen und ordnet sie in die aktuelle Forschung zur Lehrveranstaltungsevaluation ein.

2 Lehrveranstaltungsevaluation (LVE)

In der vorliegenden Arbeit wird der Begriff „Lehrveranstaltungsevaluation“ (LVE) verwendet, um in Abgrenzung zum Begriff „Lehrevaluation“ die Beschränkung auf die Evaluation konkreter Lehrveranstaltungen zu verdeutlichen. Dabei handelt es sich nur um eine Nomenklatur und nicht um eine scharfe Definition. Selbst für *Evaluation* liegen viele Definitions- und Beschreibungsversuche in der Literatur vor, die jedoch in keine einheitliche Definition münden (vgl. Abramson, 1979; Scriven, 1972; Suchman, 1967; Wittmann, 1985; Wottawa, 1986; Wottawa & Thierau, 2003). Wottawa und Thierau (2003) schlagen vor, eher die allgemeinen Kennzeichen wissenschaftlicher Evaluation herauszuarbeiten, anstatt einen weiteren Definitionsvorschlag zu liefern. Demnach besteht der Konsens darin, dass Evaluation etwas mit *Bewerten* zu tun hat. Evaluation ist *ziel-* und *zweckorientiert*. Sie soll praktische Maßnahmen *überprüfen*, *verbessern*, Entscheidungsgrundlage sein können und dem aktuellen wissenschaftlichen Stand entsprechen (vgl. Wottawa & Thierau, 2003). An diesen Kennzeichen wissenschaftlicher Evaluation orientiert sich auch die LVE.

2.1 Evaluation der Lehrveranstaltungsqualität

Die LVE hat vorrangig das Ziel, die Qualität einer Lehrveranstaltung (Lehrveranstaltungsqualität; LVQ) zu erfassen und dem Lehrenden ein Feedback über die Ergebnisse der Evaluation zu geben (vgl. Loßnitzer, Schmidt & Born, 2007; Rindermann, 2009; Schmidt & Loßnitzer, 2010; Spooren et al., 2013). Die Evaluation einer Lehrveranstaltung kann mit verschiedenen Methoden (Expertenbeurteilung, Videoanalyse, studentische Evaluation der Lehre) erfolgen. Ein auf Fragebögen basierendes Verfahren ist nur eine von vielen Möglichkeiten (Marsh & Roche, 1997). Eine weitere Möglichkeit die Qualität einer Veranstaltung zu erfassen, ist die Messung externen Kriterien. So besteht die Annahme, dass bessere Lehre auch zu erfolgreicheren Studierenden führt (Clayson, 2006). Der Lehrende mit den erfolgreicheren Studierenden ist demnach der bessere Lehrende und seine Veranstaltungen weisen eine höhere Qualität auf. Dieser Ansatz missachtet jedoch Einflussfaktoren auf die Leistung der Studierenden, die nicht dem Lehrenden zuzuschreiben sind. Operationalisiert wird die Leistung der Studenten vorrangig durch deren erreichte Note am Ende eines Semesters. Diese Note hängt al-

lerdings von sehr viel mehr Faktoren ab als lediglich von der Qualität einer Veranstaltung. Weitere Faktoren können zum Beispiel die Leistung bzw. Fähigkeit des Studierenden sein, die dieser bereits vor der Veranstaltung hat, seine individuelle Fähigkeit neues Wissen zu erwerben sowie das Interesse und die Motivation für diese Veranstaltung (vgl. Spooren et al., 2013). Auch der Wissenserwerb von Studenten wird als Maß für den Lehrerfolg herangezogen, wobei die Lehre umso besser ist, je mehr die Studenten lernen. Auch dieser Ansatz berücksichtigt nicht die individuelle Situation der Studenten. So können gute Studenten mit guten Leistungen aufgrund ihres höheren Vorwissens sehr wenig in der Veranstaltung gelernt haben, während Studenten mit sehr schlechten Vorleistungen viel Wissen erworben haben und dennoch schlechtere Leistungen zeigen als erstere (Spooren et al., 2013). Ein Rückschluss auf die LVQ oder auf die Lehrkompetenzen des Dozenten ist nur schwer möglich. Die Möglichkeiten den Einfluss der Lehrqualität unter Berücksichtigung aller relevanter Einflussgrößen zu erfassen, sind im Alltag der Universität begrenzt.

Durch die Probleme, die Qualität einer Veranstaltung über externe Kriterien zu definieren und messbar zu machen, rücken subjektive Kriterien in den Vordergrund. Eine subjektive Evaluation erfolgt meist durch die Studenten einer Veranstaltung. Mit Hilfe standardisierter Fragebögen werden verschiedene Aspekte der Lehre thematisiert und durch die Studenten bewertet. Einen sehr unspezifischen Ansatz zur Definition der LVQ im Rahmen subjektiver LVE verfolgen Spooren et al. (2013): „Teacher performance and the quality of teaching could thus be defined as the extent to which student expectations are met, thus equating student *opinions* with *knowledge*. "(Spooren et al., 2013, S.599). Mit Hilfe der LVE kann eine persönliche Meinung bezüglich verschiedener Aspekte einer Veranstaltung erhoben werden. Diese Meinungsäußerung steht bereits in Relation zu den Erwartungen der Studenten, was bei der Interpretation der LVE-Ergebnisse zu berücksichtigen ist. Welche Facetten dabei im Vordergrund stehen sollten, wird je nach Hochschule unterschiedlich betrachtet. Selbst innerhalb einer Hochschule gibt es unterschiedliche Ansichten über die Relevanz einzelner Facetten für die LVE. So sind sich Dozenten, Studenten und Evaluationsbeauftragte nicht einig, was in einen Fragebogen zur LVE aufgenommen werden soll (vgl. Spooren et al., 2013). Dennoch teilen alle Fragebögen zur LVE das Bestreben, die Qualität einer Lehrveranstaltung zu messen (vgl. Feldman, 1976; Marsh, 1983; Rindermann, 2009; Spooren et al., 2013). Durch die unterschiedlichen Fragebögen (vgl. Kapitel 2.1.1) wird deutlich, dass die Vorstellungen über die relevanten Facetten der LVQ auseinander gehen. Eine systematische Analyse der Eigenschaften, die Studierende mit guter Lehre in Verbindung bringen, liegt bereits bei Feldman (1976) vor. Er findet folgende Facetten in seiner Metaanalyse (Feldman, 1976):

- stimulation of interest
- enthusiasm
- knowledge of subject
- intellectual expansiveness
- preparation and organization
- clarity and understandableness
- elocutionary skills
- sensitivity to class level and progress
- clarity of objectives and requirements
- value of course material
- usefulness of supplementary materials
- difficulty (workload)
- fairness and evaluation
- classroom management
- feedback to students
- encouragement of discussion (openness)
- intellectual challenge
- respect for students (friendliness)
- availability and helpfulness

Diese Facetten eines Lehrenden werden in vielen Fragebögen zur Lehrveranstaltungsevaluation durch konkrete Items repräsentiert. Im internationalen Raum werden dazu sogenannte SET-Studien (Student Evaluation of Teaching) bereits seit 1927 publiziert (Clayson, 2009). Es wurde eine Vielzahl von Fragebögen zur Evaluation der Lehrqualität entwickelt, wobei vor allem der SEEQ (Students Evaluation of Educational Quality; vgl. Marsh, 1983, 1984, 1987; Marsh et al., 2009) Anwendung findet. Marsh (2007a) zeigt, welche Facetten von Feldman (1976) der SEEQ (vgl. Marsh, 1982b) abdeckt und setzt die vorhandenen Items in Beziehung zu diesen. Bei genauerer Betrachtung der Facetten von Feldman (1976) fällt auf, dass sie sich in erster Linie auf Eigenschaften des Dozenten beziehen bzw. auf von ihm beeinflussbare Größen. Möglicherweise werden dadurch Eigenschaften einer Lehrveranstaltung vernachlässigt, die ebenfalls zu einer guten Lehrveranstaltung zählen, jedoch vom Dozenten nur schwer beeinflusst werden können. Hierzu zählen zum Beispiel Rahmenbedingungen oder auch das Verhalten der Studenten. In seinem Multifaktoriellen Modell der Lehrveranstaltungsqualität führt Rindermann (2009) ähnliche Facetten wie Feldman an und erweitert das Modell guter Lehre um Aspekte der Rahmenbedingungen und Variablen der Studierenden selbst (vgl. Rindermann, 2009, S. 66). Auf Seiten der Studenten werden Facetten, wie zum Beispiel das Vorwissen und die Beteiligung der Studierenden als Einflussfaktoren auf die Qualität einer Lehrveranstaltung genannt. Die Berücksichtigung dozentenübergreifender Faktoren ist eine Stärke des multifaktoriellen Modells der LVQ. Im Allgemeinen können LVE-Fragebögen aufgrund ihrer ökonomischen Einschränkungen nicht alle Faktoren des multifaktoriellen Modells von Rindermann (2009) abdecken. Studien zur LVE basieren aus diesem Grund meist auf hochschulspezifischen Fragebögen, die unterschiedliche Facetten der LVQ erfassen wollen. Marsh (2007b) fasst in seiner Studie die Ergebnisse verschiedener Untersuchungen zur Reliabilität und Validität unterschiedlicher Fragebögen zur LVE zusammen und beschreibt, stu-

dentische Evaluationen der Lehre seien:

a) multidimensional,
b) reliabel und stabil,
c) hauptsächlich eine Funktion des Dozenten, der einen Kurs unterrichtet, als eine Funktion des Kurses der unterrichtet wird,
d) relativ valide in Bezug auf verschiedene Indikatoren für effektive Lehre,
e) relativ frei vom Einfluss oft angenommener Biasvariablen,
f) als Feedbackinstrument für Dozenten, als Mittel zur Kurswahl für Studenten und für personelle Entscheidungen nützlich.

Die Beschreibung von Marsh (2007b) wird in der Literatur von verschiedenen Autoren angezweifelt (vgl. eine Übersicht hierzu in Spooren et al., 2013). Vor allem die Validität wird kontrovers diskutiert, wobei hier vorrangig die Methode des Fragebogens als Messinstrument für die LVQ kritisiert wird. Weitere Kritikpunkte betreffen die mangelnde Qualität der Studien zur Reliabilität. Diese sind nicht auf dem Niveau, wie es zum Beispiel für Testverfahren üblich ist(vgl. Marsh et al., 2009). Vielmehr werden vorrangig exploratorische Faktorenanalysen eingesetzt und die Extraktion einer vorher postulierten Anzahl an Faktoren als Bestätigung der theoretischen Annahmen über das Messmodell interpretiert. Einig sind sich die Autoren über die Nützlichkeit der LVE als Feedbackinstrument (vgl. Abrami, d' Appolonia & Cohen, 1990; Marsh, 1982a; Marsh et al., 2009; Spooren et al., 2013). Auch die Verwendung als Mittel zur Kurswahl für Studenten und für personelle Entscheidungen auf Hochschulebene werden immer wieder genannt. Die Verwendung für personelle Entscheidungen steht jedoch im Widerspruch zu anderen Eigenschaften der LVE. Es wird nicht der Dozent evaluiert, sondern die Lehrveranstaltung eines Dozenten. Auch Selbsteinschätzungen der Studenten in Bezug auf den Kompetenzerwerb sind Teil von LVE (vgl. Braun, 2008). Rückschlüsse auf den Anteil des Kompetenzerwerbs, der dem Dozenten bzw. seiner Lehrkompetenz zuzuschreiben ist, können aufgrund der vielfältigen Einflussmöglichkeiten (vgl. hierzu Aleamoni, 1999; Spooren et al., 2013) nur schwer ermittelt werden. Die Fragebögen sind wie beschrieben mehrdimensional, sodass sich nur schwer ein isoliertes Kriterium herauskristallisieren lässt, das allein geeignet ist, die Kompetenz des Dozenten zu beschreiben.

Für den deutschsprachigen Raum haben Schmidt und Loßnitzer (2010) nach einer qualitativen und quantitativen Analyse verschiedener LVE-Fragebögen einen Definitionsvorschlag für Lehrveran-

staltungsevaluation herausgearbeitet, der den Prozess der LVE beschreibt und deren Inhalte klassifiziert.

> „Die Lehrveranstaltungsevaluation ist eine spezifische, systematische Form des lehrbezogenen Feedbacks, bei der (1) Studierende (2) schriftlich, d.h. mittels papierhafter Fragebogen oder online (3) in überwiegend standardisierter, d.h. veranstaltungs-, lehrenden- und themenübergreifender Form, (4) anhand eines strukturierten, mehrheitlich geschlossene Items/Fragen umfassenden und um einzelne offene Fragen ergänzten Erhebungsinstruments (5) Einschätzungen zu ausgewählten Aspekten des Verlaufs und der Ergebnisse einer bestimmten Lehrveranstaltung oder eines Moduls abgeben.“ (Schmidt & Loßnitzer, 2010, S.66)

Diese Einordnung beschreibt die für Schmidt und Loßnitzer (2010) wesentlichen Elemente einer LVE an Hochschulen. Es handelt sich dabei eher um eine explorative Abstraktion von Eigenschaften verschiedener Evaluationsinstrumente als um eine Definition, die logische Implikationen zur Folge hat. Für die vorliegende Arbeit liefert dieser Definitionsvorschlag eine gute Beschreibung für die LVE, die den Studien zugrunde liegt. Der formulierte Zugang zur LVE ist ökonomisch und augenscheinvalide. Der Student bewertet beobachtbares Verhalten bzw. schätzt die Ausprägung auf beobachtbaren Variablen ein, die wiederum Kennzeichen guter Lehre sind und damit LVQ im Sinne der Definition von Spooren et al. (2013) (vgl. Seite 7 der vorliegenden Arbeit) abbilden können. Die Standardisierung der LVE wird unter anderem durch die Verwendung eines einheitlichen Fragebogens erreicht.

2.1.1 Fragebögen zur LVE

Die Entwicklung von Fragebögen zur LVE orientiert sich nicht nur an den aufgeführten theoretischen Aspekten zur Erfassung der LVQ. Nahezu gleichermaßen wichtig ist die ökonomische Einbindung der Fragebögen bzw. LVE in den Universitätsalltag. Diese Anforderung zwingt die Entwickler meist, eine geringe Itemanzahl zu verwenden. Gleichzeitig soll der Fragebogen heterogen sein und vielfältige Aspekte guter Lehre abdecken (vgl. Abschnitt 2.1). Durch diese Restriktionen in der Fragebogenentwicklung unterscheiden sich Fragebögen zur Evaluation der Lehre von anderen, eignungsdiagnostischen Instrumentarien. Dennoch müssen sie sich an den gleichen Grenzwerten für Gütekriterien diagnostischer Instrumente messen lassen. Marsh (2007b) resümiert, dass die meisten Erhebungsinstrumente zur LVE eine Mischung aus logischen und pragmatischen Überlegungen darstellen, die sich gleichzeitig an psychometrischen Standards messen lassen müssen. Durch die Verwendung der LVE-Instrumente als Feedbackinstrumente, müssen dessen Items auch darauf ausgerichtet

sein, dem Lehrenden ein informatives Feedback über die Bewertung seiner Veranstaltung zu geben. Die geforderte Vielseitigkeit bzgl. der Iteminhalte bei gleichzeitig geringer Itemanzahl erschweren die Konstruktion eindimensionaler Skalen. Dennoch implizieren die Theorien zur LVQ bzw. LVE, die den Fragebögen zugrunde liegen, meist jene Eindimensionalität der verschiedenen Skalen des Instruments. Items werden nach den jeweiligen Theorien zur Lehrqualität genau einer Dimension zugeordnet. Generell werden Evaluationsinstrumente zur LVE als Feedbackinstrument konstruiert und auf ihre psychometrischen Eigenschaften hin untersucht. Man kann daher eher von einem Feedbackinstrument mit psychometrischen Eigenschaften als von einem diagnostischen Instrument zur Erhebung von Lehrqualität sprechen (Marsh, 2007b). Rindermann (2009) listet verschiedene deutschsprachige Inventare zur LVE auf und erläutert deren Dimensionalität und faktorenanalytischen Resultate in Abhängigkeit der verwendeten Analysemethode. Die Tabelle 2.1 fasst die Zusammenstellung von Rindermann (2009) und Schmidt und Loßnitzer (2010) für LVE-Fragebögen im deutschsprachigem Raum zusammen. Für eine Übersicht zu englischspachigen Fragebögen siehe Rindermann (2009) und Spooren et al. (2013). Ein einheitliches Verfahren, das deutschlandweit zur Evaluation von Lehrveranstaltungen eingesetzt wird, liegt nicht vor. Auch die inhaltliche Konstruktion und methodische Überprüfung variieren stark zwischen den unterschiedlichen Instrumenten. Durch die Heterogenität der eingesetzten Verfahren und methodischen Überprüfung wird der fehlende Standard in der LVE deutlich (vgl. hierzu auch Tabelle 2.1).

Tabelle 2.1: *Übersicht verwendeter Fragebögen zur LVE und deren faktorenanalytische Überprüfung*

Autoren	**Instrument**	**Methode**	**Ergebnisse**
Diehl und Kohr (1977), Kleine und Merkens (1979), Hofmann (1990)	Veranstaltungsbeurteilung in Psychologie (VB-Psych)	Studentenebene, Hauptkomponentenanalyse, Varimaxrotation	4 Faktoren
Müller-Wolf (1977)	Fragebogen für das Lehrverhalten in Seminaren (LVS):	Studenten- und Veranstaltungsebene, Hauptachsenmethode, Varimaxrotation	
	FA des Lehrverhaltens in Seminaren		5 Faktoren
	Einstellungen und Verhaltensreaktionen in Seminaren		3 Faktoren
	Lehrverhalten in Vorlesungen (LVV)		3 Faktoren
	Einstellungen und Verhaltensreaktionen in Vorlesungen		3 Faktoren
Winteler und Schmolck (1979, 1983)	Schätzverfahren zur Beurteilung von Lehrveranstaltungen	Studentenebene, Hauptkomponentenanalyse, Varimaxrotation	7 Faktoren
Kramis (1990)	Fragebogen zur Einschätzung der Ausbildungsqualität	Studentenebene, Hauptkomponentenanalyse, orthogonale Rotation mit Equamax	3 Faktoren
Esser (1994)	Fragen zur Veranstaltung	schiefwinklige Rotation	3 Faktoren
Basler et al. (1995)	Marburger Fragebogen zur Akzeptanz der Lehre (MFAL)	Studentenebene, Hauptkomponentenanalyse, Scree-Test, Varimaxrotation	4 Faktoren
Astleitner und Krumm (1996)	Übersetzung des Social Science Questionnaire von Murray (1983)	Konfirmatorische Faktorenanalyse	8 Faktoren
Astleitner (1991)	Übersetzung des Social Science Questionnaire von Murray (1983)	Studentenebene, Eigenwerte mind. 5% Varianzaufklärung, Hauptkomponentenanalyse, Varimaxrotation	3 Faktoren
Elbing et al. (1997)	Münchener Inventar zur Lehrveranstaltungsevaluation-Vorlesungen (MILVA-V)	Studentenebene, Hauptkomponentenanalyse, Varimaxrotation	9 Faktoren
Gold und Mayring (1997)	Ludwigsburger Rückmeldung zum Seminar	Methode unklar	5 Faktoren
Spiel und Gössler (1998)	Wiender Lehrevaluationsbogen	Studentenebene, Hauptkomponenanalyse, Eigenwerte mind. 5% Varianzaufklärung, Varimaxrotation	4 Faktoren
Westermann et al. (1998)	Fragebogen zur Beurteilung einer Lehrveranstaltung durch Studierende	Studentenebene, Hauptkomponentenanalyse, Varimaxrotation	5 Faktoren

Autoren	Instrument	Methode	Ergebnisse
Hejj (1999)	Lehrveranstaltungs-evaluationsbogen	Studentenebene, Hauptkomponentenanalyse, Scree, Varimaxrotation	3 Faktoren
Beisteiner (1999)	Eval. v. Lehrveranstaltungen an der Uni. Wien	Veranstaltungsebene, Hauptkomponentenanalyse, Scree-Test, Varimaxrotation	4 Faktoren
Staufenbiel (2000)	Fragebogen zur LVE in Vorlesungen, Seminaren und Praktika (FEVOR, FESEM, FEPRA)(Weiterentw. des VB-Psych von DiehlKohr1977a)	konfirmatorische Faktorenanalyse	4 bzw. 5 Faktoren
Multrus (1995)	veranstaltungsübergreifende Evaluation der Lehre an verschiedenen deutschen Universitäten	iterative Schätzung Kaiser-Guttmann, Varimaxrotation	6 Faktoren

Anmerkung. Die Tabelle stellt eine Zusammenfassung der Arbeiten von Rindermann (2009) und Schmidt und Loßnitzer (2010) dar und beinhaltet einen Ausschnitt von Verfahren zu denen faktorenanalytische Ergebnisse bekannt sind.

Tabelle 2.1 zeigt nur einen kleinen Ausschnitt der verfügbaren Verfahren zur LVE. Je nach Instrument werden im deutschsprachigem Raum 21 bis 66 Items zur Messung der 3 bis 21 postulierten Faktoren der Lehr- und Veranstaltungsqualität erfasst (vgl. Rindermann, 2009; Schmidt & Loßnitzer, 2010). Im Fokus der Fragebögen steht meist die Erfassung verschiedener Facetten des Lehrprozesses (Schmidt & Loßnitzer, 2010). Fragebögen wie der SEEQ (Marsh, 1982a), HILVE II (Rindermann, 2009), LeKo (Thiel, Blüthmann & Watermann, 2012), FRADOV (Koch, 2004) und der TRIL (Gollwitzer & Schlotz, 2003) sind Beispiele für etablierte Instrumente zur Erfassung des Lehrprozesses. In der neueren Entwicklung fokussieren Fragebögen stärker die Erfassung erworbener Kompetenzen der Studenten. So erfasst der BEvaKomp (Braun, 2008) verschiedene Facetten des Kompetenzerwerbs der Studierenden als Selbsteinschätzung. Auf einer Kombination aus Prozess- und Ergebnisvariablen basiert der Fragebogen PELVE (Prozess- und Ergebnisorientierte Lehrveranstaltungsevaluation; vgl. Born et al., 2006; Loßnitzer et al., 2007). Der Fragebogen PELVE wird überwiegend an der FSU Jena eingesetzt und ist Gegenstand der Untersuchungen in der vorliegenden Arbeit. Eine ausführliche Beschreibung des Fragebogens befindet sich in Kapitel 2.2.2.

2.1.2 Konfirmatorische Faktorenanalysen in der LVE

Tabelle 2.1 zeigt auf, welche Methoden zur Überprüfung der faktoriellen Struktur eingesetzt werden. Als häufigste Methode wird die Hauptkomponentenanalyse durchgeführt (vgl. Tabelle 2.1 und Rindermann, 2009). Dabei handelt es sich um eine exploratorische Faktorenanalyse (EFA) und nicht

um ein konfirmatorisches Verfahren. Die EFA ist hilfreich, um Hypothesen bezüglich der Dimensionalität der Fragebögen zu generieren. Allerdings erlaubt sie keine inferenzstatistische Prüfung einer theoretischen Faktorstruktur (für eine Gegenüberstellung von exploratorischer und konfirmatorischer Faktorenanalyse siehe z.B.: Moosbrugger & Schermeleh-Engel, 2006). Bei der EFA werden Mehrfachladungen zugelassen. Im theoretischen Messmodell wird jedes Item jedoch meistens nur genau einer Facette zugeordnet. Jede latente Variable wird durch mehrere Items konstruiert, die auf dieser latenten Variablen laden. Diese Einfachladung muss in die Modellüberprüfung einfließen, um die latenten Variablen zu konstruieren, von denen in der Theorie zur LVQ die Rede ist. Bei der Verwendung exploratorischer Faktorenanalysen zur Überprüfung der faktoriellen Struktur sind Mehrfachladungen bzw. Fehlerkorrelationen der Items enthalten und es wird die Anzahl der extrahierten Faktoren interpretiert. Die Items werden dabei dem Faktor zugeordnet, auf dem sie die höchste Ladung aufweisen. Ladungen auf anderen Faktoren werden ignoriert. Laden die intendierten Items auf dem entsprechenden Faktor, wird dies meist als Bestätigung der Theorie angesehen. Dieses Vorgehen kann jedoch nur als Hypothesen-generierendes Verfahren eingesetzt werden. Nur konfirmatorische Faktorenanalysen (CFA) ermöglichen ein theoretisch postuliertes Modell anhand einer konkreten Stichprobe zu testen. Warum Fragebögen zur LVE dennoch häufig durch exploratorische Verfahren auf ihre psychometrische Qualität und die Modellpassung überprüft werden, beschreibt Marsh et al. (2009) mit:

> „Conventional CFA goodness of fit criteria are too restrictive when applied to most multifactor rating instruments. It is my experience that it is almost impossible to get an acceptable fit (e.g., CFI, RNI, TLI > .9; RMSEA < .05) for even good multifactor rating instruments when analyses are done at the item level and there are multiple factors (e.g. 5 – 10), each measured with a reasonable number of items (e.g., at least 5 – 10 per scale) so that there are at least 50 items overall.“ (Marsh et al., 2009, S. 441)

Die wenigen konfirmatorischen Analysen legen mit der vom Modell implizierten Varianz-Kovarianz-Struktur und der zugrunde liegenden Einfachladung aller Items strenge Kriterien an die Überprüfung der Faktorstruktur des Fragebogens an. Für eine Übersicht zur Konstruktion latenter Variablen unter den Annahmen klassischer Testtheorie bzw. Item-Response-Theorie siehe Steyer, Mayer, Geiser und Cole (2015). Für die Beurteilung der Modellpassung stehen unterschiedliche deskriptive Kennwerte zur Verfügung (z. B. RMSEA, CFI, TLI; siehe hierzu Manuskript 1 in Abschnitt 4). Weiterhin kann mit inferenzstatistischen Methoden die Passung der vom Modell implizieren Varianz-Kovarianz-Struktur zur empirischen Varianz-Kovarianz-Struktur getestet werden. Oft genügen die über-

prüften Messmodelle nicht den strengen Kriterien, sodass viele derart überprüfte Fragebögen revidiert werden, indem z. B. Items ausgeschlossen oder umformuliert werden. Dennoch erreichen die oft kurzen Evaluationsinstrumente mit ihrer mehrdimensionalen Struktur nur selten die Grenzwerte für einen guten Modellfit (Marsh et al., 2009).

Eine weitere Herausforderung in der Auswahl der korrekten Methode zur Prüfung des Messmodells ergibt sich aus der Divergenz zwischen Erhebungseinheit und Analyseeinheit. Im Design der LVE evaluieren die Studenten ihre Veranstaltung und stellen damit die Erhebungseinheit in der LVE dar. Analyseeinheit ist die Veranstaltung, weil Aussagen bezüglich der Veranstaltungsqualität getroffen werden sollen (vgl. Abschnitt 2.1). Erhebungs- und Analyseeinheit unterscheiden sich demnach im Design der LVE. Dies bei der CFA zu berücksichtigen stellt eine weitere Herausforderung in der Analyse von LVE-Daten dar. In Anwendungen, bei denen die Erhebungs- und die Analyseeinheit identisch sind, wie zum Beispiel in der Eignungsdiagnostik, ist diese Thematik von geringer Bedeutung, weil man Aussagen über diejenige Person treffen möchte, die den Test oder Fragebogen bearbeitet. In Anwendungen, in denen Aussagen auf einer anderen Ebene gewünscht sind als auf der Ebene, auf der die Erhebung stattfindet bzw. die Stichprobe realisiert wird, ist eine Diskussion der Analyseeinheit notwendig. Im Rahmen der LVE handelt es sich um die Fremdeinschätzung verschiedener Aspekte der LVQ durch die evaluierenden Studenten. Die Werte der manifesten Variablen liegen somit auf Studentenebene vor, Aussagen bzgl. der Lehrqualität erfordern die Analyse auf Veranstaltungsebene. In Tabelle 2.1 ist die Analyseeinheit notiert, die zur Überprüfung der Messmodelle verschiedener Fragebögen herangezogen wurde. Mit dem Begriff *Studentenebene* ist die Analyse des unmittelbaren Resultats der Erhebung gemeint, die Rohdaten der Studenten. Dieser Ansatz kollidiert zunächst konzeptionell mit dem Anspruch der LVE-Fragebögen, Aussagen auf Ebene der Lehrveranstaltung zu treffen. Dennoch sind Werte der Studenten die häufigste Analyseeinheit. Für die Feedbackfunktion der LVE stellt das kein Problem dar. Hierfür werden je Item Verteilungskennwerte berechnet und dem Dozenten zurückgemeldet. Für die faktorenanalytischen Verfahren ist dieser Ansatz jedoch insofern problematisch, als dass die latenten Variablen auf Studentenebene definiert sind und nicht auf Veranstaltungsebene. In der Literatur wird hierzu angenommen, dass „die Verrechnung von Rohdaten [...] zumindest keine große Beeinträchtigung faktorenanalytischer Resultate zur Folge [...]“ hat (Rindermann, 2009, S. 80). Alternativ hierzu kann die CFA auch auf Basis der zuvor aggregierten Rohdaten durchgeführt werden. Hierzu werden itemspezifische Veranstaltungsmittelwerte gebildet, sodass je Veranstaltung und Item nur ein Wert vorliegt (vgl. hierzu genauer und formal Manuskript 1 in Abschnitt 4). Inwiefern sich die Ergebnisse der CFA auf Studenten- und Veranstaltungsebene

unterscheiden, untersuchten bereits Abrami (1985), Linn, Centra und Tucker (1975), Ronning und Walsh (1977), Tetenbaum (1977) und weitere. Sie kommen zu dem Schluss, dass die faktorielle Struktur relativ robust bei der Analyse der Rohdaten (Analyse auf Studentenebene) ist. Im deutschsprachigem Raum werden die Analysen sowohl auf Studentenebene (Rohdatenanalyse) als auch auf Veranstaltungsebene (Veranstaltungsmittelwerte) durchgeführt (vgl. Tabelle 2.1). Für viele Autoren ist die Analyse der Veranstaltungsmittelwerte und damit die Analyse auf Veranstaltungsebene eine Möglichkeit, potentiell störende Einflüsse auf individueller Ebene zu beseitigen und gleichzeitig Aussagen auf Veranstaltungsebene treffen zu können (vgl. hierzu Rindermann, 2009). Im internationalen Raum werden hauptsächlich die aggregierten Variablen analysiert (vgl. Marsh, 2007a). Clayson (2007) betont diesen Aspekt und behauptet: „Currently, researchers, such as Marsh and Roche, disallow any research that would use within-class data [...]“ (Clayson, 2007, S. 35).

Aufgrund vielfältiger Ansätze und dem geringen Konsens bezüglich der zu verwendeten Analyseebene, empfiehlt Rindermann (2009) verschiedene Methoden der CFA und verschiedene Analyseeinheiten (Veranstaltungsmittelwerte vs. Rohdatenanalyse) vergleichend gegenüber zu stellen. Studie 1 (vgl. Manuskript 1 in Abschnitt 4) geht genauer auf die Problematik der Analyseeinheit ein und vergleicht verschiedene CFA-Verfahren. Studie 2 (vgl. Manuskript 2 in Abschnitt 5) fokussiert hingegen den Vergleich verschiedener Stichprobentypen. Die Erkenntnisse aus beiden Studien fließen in die Treatmentstudie ein (vgl. Manuskript 3 in Abschnitt 6).

2.1.3 Stichproben in der LVE

Abbildung 2.1 skizziert die Zuordnung von Veranstaltungen zu Dozenten und von Studenten zu Veranstaltungen. Die Skizze verdeutlicht, wie durch die Zusammenführung verschiedener veranstaltungsspezifischer Stichproben neue strukturelle Eigenschaften in der resultierenden Gesamtstichprobe auftreten können. Studenten evaluieren unterschiedliche Veranstaltungen desselben oder verschiedener Dozenten. In einer derartigen Gesamtstichprobe liegt Mehrfachevaluation auf Studentenebene vor. Sofern unterschiedliche Veranstaltungen desselben Dozenten enthalten sind, ist auch auf Veranstaltungsebene Mehrfachevaluation enthalten. Nach dem Definitionsvorschlag für die LVE (vgl. Abschnitt 2.1), evaluieren Studenten verschiedene Aspekte der Lehrveranstaltungsqualität. Evaluationsergebnisse einer LVE sollen auf die Qualität der Veranstaltung zurückgeführt werden können. Sind in der Stichprobe mehrere Evaluationen derselben Studenten enthalten, kann nicht mehr davon ausgegangen werden, dass die Ergebnisse dieser Veranstaltung stochastisch unabhängig voneinander sind. Die Annahme unabhängiger und identisch verteilter Zufallsvariablen (iid-Annahme;

independently and identically distributet; vgl. Skinner et al., 1989) ist damit verletzt. Die Ergebnisse und Zusammenhänge verschiedener Veranstaltungen, die (teilweise) dieselben Studenten beinhalten (Mehrfachevaluation auf Studentenebene), hängen demnach nicht nur von den Aspekten der Lehrveranstaltung, sondern auch den personenspezifischen Eigenschaften ab.

Innerhalb einer Veranstaltung werden Mehrfachevaluationen per Design ausgeschlossen. Damit ist die iid-Annahme für die Beobachtungen einer Veranstaltung plausibel (sofern ein Student genau einen Evaluationsbogen ausfüllt). Für die Feedbackfunktion der LVE, basierend auf der veranstaltungsspezifischen Auswertung, sind Mehrfachevaluationen demnach nicht von Bedeutung. Nach der Zusammenführung veranstaltungsspezifischer Stichproben kann die resultierende Gesamtstichprobe Mehrfachevaluationen beinhalten. Abbildung 2.1 suggeriert eine Identifizierbarkeit der Studenten, sodass die Mehrfachevaluation beobachtbar und nachträglich korrigierbar erscheint. Allerdings ist häufig unbekannt, wer genau die Veranstaltungen evaluiert. Ohne eine genaue Identifikation der Studenten bzw. Beobachtungen von gleichen Studenten, kann der Effekt der Mehrfachevaluation auf Ergebnisse der Modellprüfung nicht untersucht werden. In diesem Fall können Mehrfachevaluationen bei einer CFA auch nicht berücksichtigt werden. Inwiefern die Mehrfachevaluation auf Studentenebene bei einer CFA von Bedeutung ist, wird in Studie 2 (vgl. Manuskript 2 in Abschnitt 5) näher erläutert und empirisch untersucht.

Abbildung 2.1 veranschaulicht, dass, nach Zusammenführung mehrerer veranstaltungsspezifischer Stichproben, die Mehrfachevaluation auch auf Dozentenebene in der Gesamtstichprobe enthalten sein kann. LVE-Ergebnisse von Veranstaltungen desselben Dozenten können stärkere Zusammenhänge aufweisen als LVE-Ergebnisse von Veranstaltungen unterschiedlicher Dozenten. Auch in diesem Fall hätte die Mehrfachevaluation auf Dozentenebene die Verletzung der iid-Annahme zur Folge und kann die CFA-Ergebnisse beeinflussen. Veranstaltungen desselben Dozenten zeigen ähnlichere Ergebnisse als Veranstaltungen unterschiedlicher Dozenten, auch unabhängig vom Inhalt der Lehrveranstaltung (vgl. Marsh, 1987, 2007a).

Müsste man ein Design zur Untersuchung der faktoriellen Struktur des Fragebogens wählen, würde man sich für ein Design entscheiden, indem die Mehrfachevaluation auf Dozenten- oder Studentenebene nicht enthalten ist. Toland und de Ayala (2005) haben in ihrer Studie auf Studentenebene nur eine Evaluation und auf Dozentenebene nur eine Veranstaltung zugelassen und eine Stichprobe zur Untersuchung der faktoriellen Struktur ihres Fragebogens direkt und ohne Zusammenführung verschiedener LVE erhoben. Dieses Vorgehen vermeidet Mehrfachevaluation, ist jedoch unüblich und nur aufwendig im Rahmen der alltäglichen LVE umsetzbar. Häufiger werden veranstal-

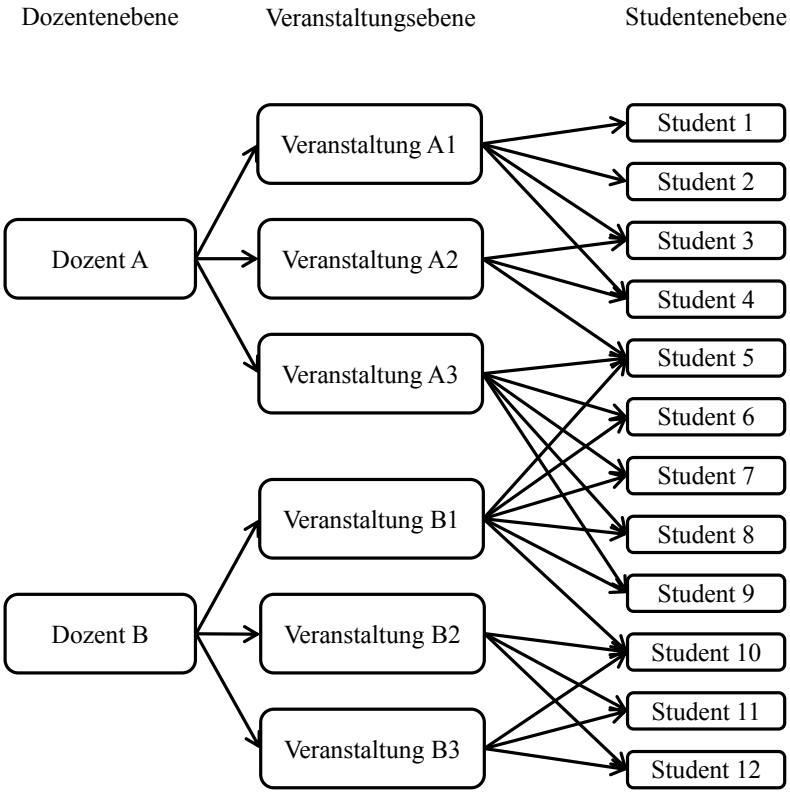


Abbildung 2.1: Stichprobenzusammensetzung in der Lehrveranstaltungsevaluation

tungsspezifische Stichproben zu Gesamtstichproben zusammengeführt und die resultierenden strukturellen Probleme der Daten nicht weiter berücksichtigt. Sofern dieser Sachverhalt bei der Analyse einer konkreten Stichprobe unberücksichtigt bleibt, kann die Mehrfachevaluation auf Dozentenebene und auf Studentenebene einen Einfluss auf die Ergebnisse verschiedener Analysen haben (vgl. Skinner et al., 1989). Gegenstand der vorliegenden Arbeit ist eine empirische Untersuchung des Einflusses dieser Mehrfachevaluation auf die Ergebnisse der verschiedenen Verfahren zur konfirmatorischen Faktorenanalyse (vgl. Manuskript 2 in Abschnitt 5). Zum Großteil werden CFA zur Überprüfung der Modellpassung von LVE-Fragebögen auf Veranstaltungsebene unter Verwendung von Veranstaltungsaggregaten durchgeführt. Dabei wird angenommen, dass die stochastischen Abhängigkeiten zwischen verschiedenen Lehrveranstaltungsevaluationen mit teilweise den gleichen Studenten (Mehrfachevaluation auf Studentenebene) beseitigt sind und kein Einfluss hierdurch auf die Kriterien der Modellpassung beobachtbar ist (vgl. Rindermann, 2009). Zum Teil werden diese Analysen auf Veranstaltungsebene durchgeführt, weil keine personenbezogenen Informationen über die Studenten vorliegen. Eine empirische Untersuchung des Einflusses dieser Mehrfachevaluation auf die Modellpassung des postulierten Messmodells liegt bisher im Rahmen der Lehrveranstaltungsevalua-

tion noch nicht vor. Rindermann (2009) schlägt vor, „[...] verschiedene Methoden und Datensätze in ihren Resultaten einander vergleichend gegenüber zu stellen“ (Rindermann, 2009, S.82), jedoch wird nicht erläutert, wie mit den Ergebnissen zu verfahren ist. Wie sollten unterschiedliche Ergebnisse bzgl. der Modellpassung bewertet werden? Welches Verfahren ist dann besser geeignet? In Studie 2 wird der geforderte Methodenvergleich und der Vergleich verschiedener Stichprobentypen durchgeführt. Hierfür (vgl. Manuskript 2 in Abschnitt 5) werden verschiedene Stichproben betrachtet, die in unterschiedlichem Ausmaß die iid-Annahme verletzten. Es werden vier verschiedene Stichprobentypen den CFA zugrunde gelegt, die sich durch *Dozenten- und Studentenwiederholung, nur Dozentenwiederholung, nur Studentenwiederholung* oder *weder Dozenten- noch Studentenwiederholung* auszeichnen. Ob die Mehrfachevaluation einen Einfluss auf die Resultate der CFA hat, wird in Studie 2 für verschiedene CFA-Verfahren (siehe Manuskript 1 in Abschnitt 4) untersucht.

2.2 Lehrveranstaltungsevaluation an der FSU Jena

An der FSU Jena wird seit 1997 die LVE durch das Universitätsprojekt Lehrevaluation (ULe) durchgeführt und weiterentwickelt. Das Portfolio des ULe umfasst verschiedene Befragungen für Lehrveranstaltungen und Studiengänge. Studieneingangsbefragungen, Zwischenbilanzen und Studienabschlussbefragungen sind als Instrumente der Qualitätssicherung auf Studiengangsebene verankert. Die LVE soll ebenfalls als Instrument zur Qualitätssicherung der Lehre beitragen. Sie hat den Anspruch, die Qualität von Lehrveranstaltungen zu erfassen und die Qualitätsentwicklung in der Lehre zu fördern (vgl. Vetterlein & Sengewald, 2015). Hierfür wurde die LVE institutionalisiert und in der Evaluationsordnung verankert (vgl. Friedrich-Schiller-Universität Jena, 2012). Sie ist explizit nicht als Steuerungselement im Evaluationsportfolio des ULe integriert, sondern als Feedbackinstrument für den Lehrenden der Veranstaltung. Zentral ist die weit gefasste Regelung zur Durchführung der LVE, die sich an den Interessen der Dozenten orientiert. Dozenten können demnach selbst entscheiden, ob und wie sie ihre Lehrveranstaltung evaluieren oder evaluieren lassen. Sie haben jederzeit die Möglichkeit, sich an das ULe zu wenden und die Evaluation ihrer Lehrveranstaltung in Auftrag zu geben. Die Evaluation ist damit ein Prozess, der durch den Dozenten freiwillig initiiert und bis zum Ergebnisbericht von ULe gesteuert wird.

2.2.1 Ablauf einer LVE

Nachdem der Dozent die LVE seiner Veranstaltung bei ULe in Auftrag gegeben hat, wird ein hoch standardisierter Prozess in Gang gesetzt. Jeder Dozent, der mit ULe evaluiert, hat dort ein Benutzerkonto, indem er neue Veranstaltungen zur Evaluation anmelden kann und die Evaluationsergebnisse vergangener LVE wiederfindet. Abbildung 2.2 zeigt schematisch den Ablauf einer LVE.

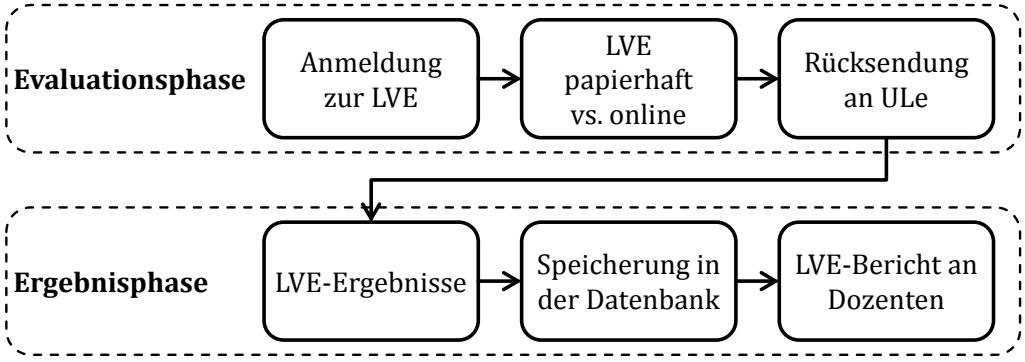


Abbildung 2.2: Schematische Darstellung des Ablaufs einer Lehrveranstaltungsevaluation

Der Dozent entscheidet selbst, welche Veranstaltung er wann und wie evaluieren lässt. Grundsätzlich empfiehlt das ULe eine Evaluation in der Mitte des Semesters, um ggf. Anpassungen der eigenen Lehre noch für die evaluierte Veranstaltung umzusetzen. Erfahrungsgemäß evaluieren die Dozenten jedoch vorrangig gegen Ende des Semesters. Eine zügige Rückmeldung der Evaluationsergebnisse ermöglicht es dem Dozenten, die Ergebnisse mit den Studenten der Veranstaltung zu besprechen. Bei der Anmeldung zur Evaluation wählt der Dozent eine passende Methode (Online vs. Papier) und eine Fragebogenversion. Es kann zwischen Fragebögen für Vorlesungen, Seminare und Übungen gewählt werden. Die Fragebögen sind im Anhang A dieser Arbeit einsehbar. Eine genaue Beschreibung der enthaltenen Items befindet sich in Kapitel 2.2.2. Nach der erfolgreichen Anmeldung der Lehrveranstaltung zur Evaluation werden die Materialien an den Dozenten versendet. Der Dozent erhält die Fragebögen für die Studierenden und einen Dozentenfragebogen, indem er die Veranstaltung aus seiner eigenen Perspektive bewertet. Die papierhafte Durchführung der LVE erfolgt meist in einer konkreten Veranstaltung des Dozenten durch die anwesenden Studenten. Im Anschluss sendet der Dozent die ausgefüllten Fragebögen zurück an das ULe. Es schließt sich die Ergebnisphase an (vgl. Abbildung 2.2). Zunächst werden die Fragebögen eingescannt und automatisiert durch das ULe-Tool verarbeitet. Die Rohdaten werden in einer Datenbank gespeichert und für den Ergebnisbericht entsprechend auf Veranstaltungsebene durch Angabe des Mittelwerts und weiterer Verteilungsparameter grafisch und tabellarisch aufbereitet. Im Anschluss werden verschiedenene Ergebnisdokumente (Evaluationsbericht, Präsentation und Aushang) an den Dozenten versendet und im Benutzerkon-

to archiviert. Wie Dozenten mit den Ergebnissen ihrer LVE umgehen, ob sie die Berichte lesen und die Ergebnisse mit ihren Studenten diskutieren und Veränderungsmaßnahmen ableiten, liegt ausschließlich in der Verantwortung der Dozenten selbst. Der Umgang des Dozenten mit den Ergebnissen der LVE ist weitestgehend unbekannt. Studie 3 (vgl. Manuskript 3 in Abschnitt 6) geht der Frage nach, ob die Ergebnisberichte einen Einfluss auf die Qualität nachfolgender Lehrveranstaltungen haben können. Hierfür wurde eine Treatmentstudie durchgeführt, in der jeder Dozent einen von zwei unterschiedlichen Ergebnisberichten erhielt. Im Kapitel 2.2.3 wird diese Problematik näher erläutert und es werden Vorstudien vorgestellt, die zunächst den Umgang mit den LVE Ergebnissen untersuchen.

Durch die fortlaufende Speicherung der Evaluationsergebnisse in der Datenbank und die Archivierung der Ergebnisberichte kann der Dozent jederzeit auch ältere LVE-Ergebnisse einsehen. Der Dozent hat zudem die Möglichkeit, eine oder mehrere seiner Veranstaltungen des laufenden Semesters evaluieren zu lassen. Im Laufe der Zeit liegen damit möglicherweise mehrere Evaluationen eines Dozenten vor. Fasst man sie in einem Datensatz zusammen, liegt eine Gesamtstichprobe vor, die Mehrfachevaluation auf Dozentenebene enthält.

Auch auf Studentenebene ist die Evaluation freiwillig. Lässt ein Dozent seine Veranstaltung evaluieren, ist damit nicht sicher, dass auch alle Studenten den Fragebogen ausfüllen. Der Evaluationszeitpunkt ist für alle Studierenden im Falle der papierhaften Evaluation identisch oder im Falle der Online-Evaluation individuell verschieden. Jede Evaluation (ausgefüllter Fragebogen) wird bei der Berechnung der Item-Mittelwerte berücksichtigt. Es ist möglich, dass ein Student im Laufe seines Studiums mehrere Veranstaltungen evaluiert. Im Evaluationsbogen des ULe wird ein persönlicher Code vom Studenten erfragt, sodass eine Mehrfachevaluation auf Studentenebene identifizierbar ist. Bei dem Personencode handelt es sich um anonymisierte persönliche Informationen. Der Personencode wird aus dem ersten Buchstaben des Geburtsorts, dem zweiten Buchstaben des Vornamens und dem dritten Buchstaben des Nachnamens (Geburtsnamens) sowie den jeweils letzten Ziffern des Geburtstags und Geburtsmonats gebildet.

2.2.2 Der Fragebogen PELVE an der FSU Jena

Der Fragebogen zur Prozess- und Ergebnisorientierten Lehrveranstaltungsevaluation (PELVE; Born et al., 2006; Loßnitzer et al., 2007) wurde an der FSU Jena entwickelt. Er erfüllt die Kriterien der Definition der Lehrveranstaltungsevaluation (siehe Seite 10) und vereint Prozess- und Ergebnisvariablen in einem Fragebogen (vgl. Tabelle 2.2). Seit der Gründung des Universitätsprojekts Lehr-

evaluation 1997 durchlief der PELVE mehrere Entwicklungsphasen. Aufbauend auf dem Qualitätsmodell von Buhl (1999) wurden zunächst einzelne Items für die verschiedenen theoretischen Facetten der LVQ entwickelt (vgl. Loßnitzer et al., 2007). Darauf aufbauend wurden nach mehreren Revisionsphasen im Sommersemester 2004 drei Versionen des PELVE in den Regelbetrieb aufgenommen Fragebogen für Seminare, Vorlesungen, Übungen; vgl. Loßnitzer et al. (2007). Seit dem Sommersemester 2005 wurde die Dokumentation und die Speicherung der Evaluationsdaten vereinheitlicht, sodass die Analysen der Studien in der vorliegenden Arbeit auf eine breiter Datenbasis mit allen LVE seit dem Sommersemester 2005 beruhen. Im Folgenden wird die Konzeption und die Entwicklung des Fragebogens kurz dargestellt.

2.2.2.1 Konzeption des Fragebogens

Der Fragebogen PELVE wurde in Anlehnung an das Qualitätsmodell des ULe entwickelt (vgl. Buhl, 1999; Loßnitzer et al., 2007). Das folgende Zitat zeigt, was unter der Qualität einer Lehrveranstaltung in der vorliegenden Arbeit verstanden wird.

> „Das Qualitätsmodell des Universitätsprojekts Lehrevaluation definiert die Qualität einer Lehrveranstaltung multidimensional als Zusammenspiel von unmittelbar oder mittelbar durch die Veranstaltungsteilnehmenden veränderbaren Prozessvariablen einerseits und den zu erreichenden bzw. erreichten Ergebnissen der Lehrveranstaltung andererseits“ (Loßnitzer et al., 2007, S. 328).

Tabelle 2.2: *Qualitätsmodell der Lehre an der FSU Jena*

Qualität einer Lehrveranstaltung	
Zusammenwirken von Prozess- und Ergebnisvariablen	
Prozessvariablen	**Ergebnisvariablen**
Prinzipiell veränderbar, bilden den Prozess der Lehrveranstaltung ab	Veranstaltungsspezifisch, bilden Ergebnisse der Lehrveranstaltung ab
Rahmenbedingungen (Raum, Ausstattung, Lärm)	Gesamteindruck (Attraktivität der Lehrveranstaltung)
Verhalten des Dozenten (Aufbereitung und Darstellung des Stoffes, Leitungsfunktion)	Zuwachs an Wissen, Fertigkeiten, Kompetenzen (Fachwissen, praktische Anwendung, Fachübergreifendes Denken)
Verhalten der Studierenden (Teilnahmeregelmäßigkeit, Aufmerksamkeit, Lerneinsatz)	

Anmerkung. Tabelle entnommen aus Born et al. (2006)

Eine genauere Ausdifferenzierung der Prozess- und Ergebnisvariablen nehmen Born et al. (2006) vor (vgl. Tabelle 2.2). Dabei werden entsprechend des Multifaktoriellen Modells der Lehrveranstaltungsqualität nach Rindermann (2009) sowohl Variablen zum Verhalten des Dozenten als auch Rahmenbedingungen und Studentenverhalten durch den Fragebogen abgedeckt. Die Beurteilung der Ergebnisse einer Lehrveranstaltung umfasst eine Gesamteinschätzung als globales Maß der Zufriedenheit und eine Selbsteinschätzung bzgl. des Kompetenzerwerbs. An dieser Stelle sei erwähnt, dass es sich bei der Verwendung der Begriffe Ergebnisvariablen und Prozessvariablen nicht um Variablen im engeren statistischen Sinne handelt, die klar definiert sind. Es sind weder Zufallsvariablen noch Variablen einer konkreten Stichprobe, die Werte annehmen können. Sie repräsentieren vielmehr eine Kategorisierung verschiedener theoretischer Größen im Rahmen der Lehrveranstaltungsqualität. So sind je nach Fragebogen unterschiedliche Facetten und Items unter dem Begriff Prozessvariablen verortet. Für die fünf Bereiche aus Tabelle 2.2 wurden Items entwickelt, die den entsprechenden Inhaltsbereich abdecken sollen. Dabei fokussieren die Prozessvariablen Aspekte einer Lehrveranstaltung, die prinzipiell veränderbar sind und sich auf den Prozess der Lehrveranstaltung beziehen. Ergebnisvariablen des PELVE-Fragebogens decken in diesem Fall zwei mögliche Facetten ab, die als Ergebnis einer Lehrveranstaltung verstanden werden können. Diese unterteilen sich in *Gesamteindruck* und *Kompetenzerwerb*. Der Gesamteindruck soll durch Items zur Zufriedenheit mit der Lehrveranstaltung erhoben werden. Der selbsteingeschätzte Kompetenzerwerb soll hingegen durch Items über den Zuwachs an Wissen, Fertigkeiten und Kompetenzen erhoben werden. Der selbsteingeschätzte Kompetenzerwerb lässt dabei keine Rückschlüsse auf das Kompetenzniveau der Studenten einer Veranstaltung zu. Auch der tatsächliche Zuwachs an Wissen und Kompetenzen wird damit nicht gemessen. Welche Subgruppen der Studenten einer Veranstaltung angeben viel bzw. wenig gelernt zu haben, kann quantitativ auf Basis der Antwortverteilung abgeschätzt werden. Die qualitative Bewertung kann nur in der Diskussion mit den Studenten erfolgen und sollte in Zusammenhang mit den Ergebnissen der anderen Facetten des PELVE-Fragebogens interpretiert werden.

Die drei Fragebogenversionen (Vorlesung, Seminar, Übung) beinhalten 35 Items zur Erfassung der fünf Facetten zu den Prozess- und Ergebnisvariablen. Darüber hinaus unterscheiden sich die drei Fragebogenversionen aufgrund zusätzlicher, spezifischer Items. Der Seminarfragebogen enthält acht, der Vorlesungsfragebogen vier und die Version für Übungen sieben spezifische Items. Die Fragebögen sind als Anhang A beigefügt. In den eingebunden Studien wird das Messmodell des PELVE-Fragebogens untersucht bzw. angewendet. Abgeleitet wird dieses Messmodell aus der Theorie zur LVQ, die der Fragebogenkonstruktion zugrunde lag und den Erkenntnissen aus der Arbeit von Born et al. (2006),

Tabelle 2.3: *Zuordnung der 35 Items des PELVE-Fragebogens zu den Bewertungsdimensionen*

Postulierte Bewertungsdimension - Unabhängig von der Veranstaltungsform			
Bewertungsdimension	Itemanzahl	Itembezeichnung	Beispiel-Item
Gesamteindruck (sal)	5 Items 1 Gesamtitem	sal_1,..., sal_5 sal_6	sal_3: Die Veranstaltung versetzt mich in die Lage, die Inhalte selbstständig zu vertiefen.
Kompetenzerwerb (squ)	8 Items 1 Gesamtitem	squ_1,...,squ_8 squ_9	squ_3: Wissen über Forschungsverfahren und wissenschaftliche Methoden
Rahmenbedingungen (sra)	5 Items 1 Gesamtitem	sra_1,...,sra_5 sra_6	sra_3: Die Veranstaltung findet in einem angemessenen zeitlichen Rahmen (Zeitpunkt, Dauer, Überschneidungen, ...) statt.
Dozentenverhalten (sdo)	8 Items 1 Gesamtitem	sdo_1,...,sdo_8 sdo_9	sdo_3: teilt die Veranstaltungszeit sinnvoll ein (auf Vortrag, Diskussion, Klärung von Fragen, ...).
Studentenverhalten (ste)	4 Items 1 Gesamtitem	ste_1,..., ste_4 ste_5	ste_3: beteiligen sich, soweit möglich, aktiv an der Veranstaltung.

Anmerkungen. Berücksichtigt werden hier nur die Items, die für jede Veranstaltungsform eingesetzt werden. Die Abkürzungen der Bewertungsdimensionen und Itembezeichnungen werden im Messmodell verwendet. Eine vollständige Liste aller Items, die in das Messmodell des PELVE eingehen, befindet sich in Tabelle 2.4

die ein seminarspezifisches Messmodell geprüft haben. Als Ausgangspunkt werden die 35 Items in die Betrachtung einbezogen, die in allen Versionen des PELVE enthalten sind und die für ein veranstaltungsübergreifendes Messmodell herangezogen werden. In Anlehnung an Born et al. (2006) zeigt Tabelle 2.3 die Anzahl der Items für jede postulierte Bewertungsdimension. Zu jeder dieser Dimensionen existiert ein Gesamtitem, welches eine abschließende Gesamtbeurteilung zur entsprechenden Bewertungsdimension erhebt. Darüber hinaus ist jeweils ein Beispielitem gegeben. Alle Items, die später im Messmodell Verwendung finden, sind in Tabelle 2.4 enthalten.

2.2.2.2 Entwicklung eines Messmodells des PELVE-Fragebogens

Aus der Zuordnung der Items zu den Bewertungsdimensionen des PELVE (vgl. Tabelle 2.3) lässt sich ein Messmodell mit fünf Dimensionen konstruieren. Jede dieser Dimensionen wird durch die entsprechenden Items gemessen. In Abbildung 2.3 ist das theoriegeleitete Messmodell schematisch dargestellt. Auf die Darstellung jedes einzelnen Items wird aus Platzgründen verzichtet, sodass nur

jeweils das erste und das letzte Item der entsprechenden Bewertungsdimension abgebildet sind. Die Theorie macht zunächst keine konkreten Aussagen über die Korrelation der latenten Variablen. In der hier verwendeten Definition der Lehrveranstaltungsqualität (vgl. Kapitel 2.2.2.1) ist jedoch vom Zusammenspiel der Prozess- und Ergebnisvariablen die Rede. Korrelationen zwischen den latenten Variablen sollten demnach im Modell zugelassen werden.

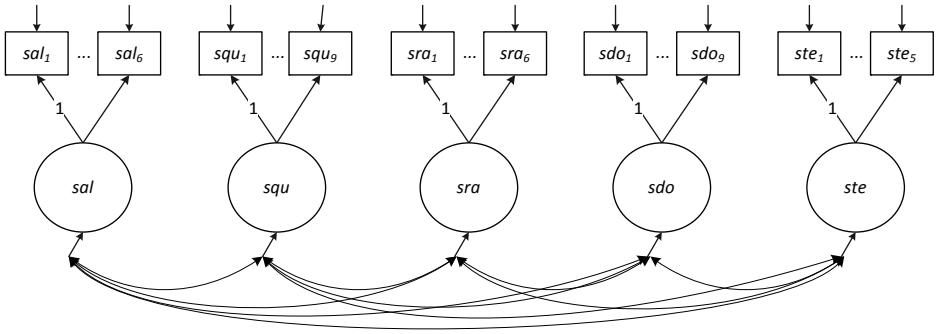


Abbildung 2.3: Schematische Darstellung des theoretischen Messmodells des PELVE
Anmerkungen. Es ist das jeweils erste und letzte Item der theoretischen Facette dargestellt und entsprechend der Variablenbezeichnung benannt. Variablenbezeichnungen: sal (Gesamteindruck); squ (Kompetenzerwerb); sra (Rahmenbedingungen); sdo (Dozentenverhalten); ste (Studentenverhalten)

In der Arbeit von Born et al. (2006) wird dieses Modell überprüft. Dafür werden $N_s = 2774$ ausgefüllte Evaluationsbögen auf Studentenebene verwendet, die insgesamt $N_v = 367$ Veranstaltungen zugeordnet sind (vgl. Born et al., 2006). Bei diesen Veranstaltungen handelt es sich ausschließlich um Seminare. Vorlesungen und Übungen werden nicht berücksichtigt. Aus diesem Grund verwenden Born et al. (2006) die zusätzlichen Items des Fragebogens für Seminare und konstruieren eine weitere latente Variable hierfür (vgl. Abbildung 2.4). Darüber hinaus wird eine theoriegeleitete Trennung der Dimension *Kompetenzerwerb* vorgenommen. Das Modell von Born et al. (2006) beinhaltet demnach die latenten Bewertungsdimensionen *Gesamteindruck*, *Fachkompetenz*, *sonstige Kompetenzen*, *Rahmenbedingungen*, *Dozentenverhalten*, *Dozentenverhalten (seminarspezifisch)*, *Studentenverhalten* und *Studentenverhalten (seminarspezifisch)*. Die Benennung der latenten Variablen spiegelt in der LVE nicht immer den exakten Inhalt wider, sodass bei genauerer Betrachtung der Items auch Zweifel an der Korrektheit der Benennung der latenten Variablen aufkommen könnten. So ist zum Beispiel unter der Benennung Dozentenverhalten nicht zu verstehen, dass die hier ein metrischer Wert des Ausmaßes an Verhalten des Dozenten gemessen wird. Vielmehr handelt es sich um die Qualität des Dozentenverhalten, die quantifiziert werden soll. Dennoch folgt die vorliegende Arbeit der Nomenklatur der Autoren des PELVE-Fragebogens (Loßnitzer et al., 2007) und benennt die verschiedenen Bewertungsdimensionen entsprechend, um in der Terminologie der LVE zu bleiben. Welche Items den Bewertungsdimensionen zugeordnet werden, kann der Tabelle 2.4 entnommen werden.

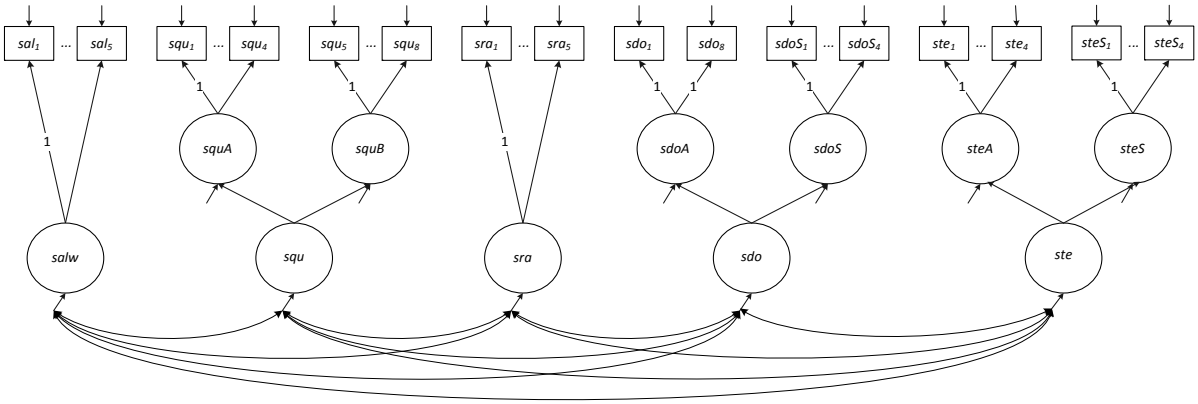


Abbildung 2.4: Messmodell des PELVE-Fragebogens nach Born et al. (2006)
Anmerkungen. Es wurden die folgenden acht latenten Variablen erster Ordnung und drei latenten Variablen zweiter Ordnung konstruiert: sal (Gesamteindruck); $squA$ (Fachkompetenz); $squB$ (sonstige Kompetenzen); squ (Kompetenzerwerb); sra (Rahmenbedingungen); $sdoA$ (allgemeines Dozentenverhalten); $sdoS$ (seminarspezifisches Dozentenverhalten); sdo (Dozentenverhalten); $steA$ (allgemeines Studentenverhalten); $steS$ (seminarspezifisches Studentenverhalten); ste (Studentenverhalten)

Die Überprüfung des Messmodells aus Abbildung 2.4 wurde von Born et al. (2006) auf Studentenebene ohne explizite Berücksichtigung potentieller Mehrfachevaluationen durchgeführt. Die Autoren geben an, mit Hilfe einer konfirmatorischen Faktorenanalyse die Fünf-Faktorenstruktur des Fragebogens beibehalten zu können (χ^2 = 6707.29 df= 652, p = .000; RMSEA= .058). Diese Zählung berücksichtigt nicht die Unterteilung in Faktoren erster und zweiter Ordnung. Abbildung 2.4 zeigt, dass acht Faktoren erster Ordnung in diesem Modell spezifiziert sind. Zwei dieser Faktoren ($steS$ und $sdoS$) repräsentieren seminarspezifische Aspekte des Studenten- bzw. Dozentenverhaltens und können nicht mit dem veranstaltungsübergreifenden Qualitätsmodell der Lehre vereinbart werden. Die übrigen sechs latenten Variablen können unabhängig vom Veranstaltungstyp konstruiert werden. Hierbei erfolgt eine Unterteilung der Kompetenzdimension in fachspezifische Kompetenzen ($squA$) und sonstige Kompetenzen ($squB$). Diese sogenannten *sonstigen Kompetenzen* stellen eine Mischung aus Methoden-, Sozial- und Personalkompetenz dar (vgl. Born et al., 2006). Die Spezifikation einer latenten Variable zweiter Ordnung (squ) mit freien Ladungen auf die Subdimension scheint aus theoretischer Perspektive nicht notwendig, weil es von der Theorie nicht gefordert wird. Ein wesentlicher Kritikpunkt an dem dargestellten Modell ist der fehlende Bezug der latenten Variablen zur Theorie über die Qualität von Lehrveranstaltungen. Die latenten Dimensionen sind aufgrund der Analyseebene (Analyse der Rohdaten auf Studentenebene) nur auf Studentenebene definiert und demnach sind die Werte der latenten Variablen als Werte der Studenten auf der jeweiligen Dimension zu interpretieren. Der Rückschluss auf die Qualität der Lehrveranstaltung, die laut Theorie durch den Fragebogen erhoben werden soll, erfolgt nicht. Darüber hinaus werden die Gesamt-

items (vgl. Tabelle 2.3) der Bewertungsdimensionen nicht berücksichtigt (vgl. Born et al., 2006). Das Messmodell in Abbildung 2.4 beinhaltet demnach nur 30 der 35 Kernitems, die unabhängig vom Veranstaltungstyp die Lehrqualität erfassen sollen. Als Ausblick empfehlen die Autoren eine Teilung der Bewertungsdimension *Rahmenbedingungen* in zwei Subdimensionen (Born et al., 2006).

Die vorliegende Arbeit greift die Ergebnisse der Studie auf und konstruiert sieben latente Variablen erster Ordnung, die durch entsprechende Items der einzelnen Bewertungsdimensionen gemessen werden sollen. Dabei werden auch die Gesamtitems berücksichtigt, sofern sich diese eindeutig einer Bewertungsdimension zuordnen lassen. Durch die Unterteilung der Bewertungsdimension *Kompetenzerwerb* und *Rahmenbedingungen* ist hier keine Zuordnung des Gesamtitems möglich. Das Messmodell (vgl. schematisch Abbildung 2.5) wird unabhängig vom Veranstaltungstyp spezifiziert, sodass die seminar-, vorlesungs- und übungsspezifischen Items nicht berücksichtigt werden.

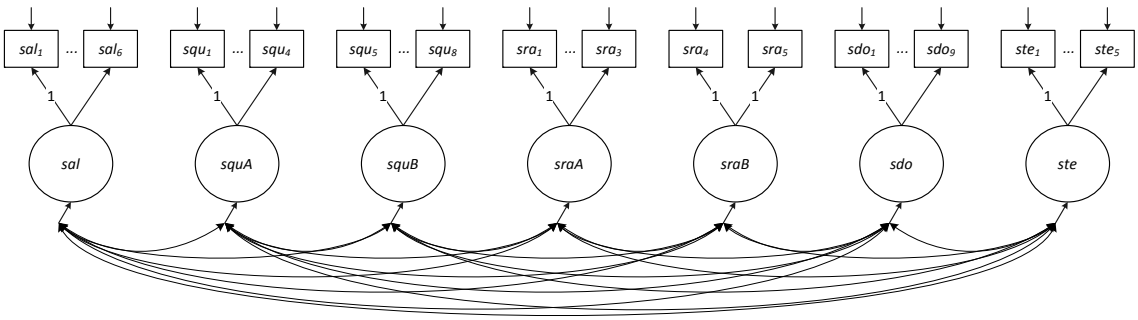


Abbildung 2.5: Schematische Darstellung des postulierten PELVE-Messmodells auf Studentenebene
Anmerkungen. Die Abbildung zeigt das theoretische Messmodell in Anlehnung an die Ergebnisse von Born et al. (2006). Variablenbezeichnung: sal (Gesamteindruck); $squA$ (Fachkompetenz); $squB$ (sonstige Kompetenzen); $sraA$ (Rahmenbedingungen); $sraB$ (Begleitmaterialien); sdo (Dozentenverhalten); ste (Studentenverhalten). Auf die Darstellung aller Items der entsprechenden latenten Variablen wird verzichtet. Sie sind in Tabelle 2.4 enthalten.

Zusammenfassend kann durch die Arbeit von Born et al. (2006) eine empirische Unterstützung der theoretischen Bewertungsdimensionen gezeigt werden. Die Studien in der vorliegenden Arbeit bauen auf den Ergebnissen von Born et al. (2006) auf, integrieren die Gesamtitems und fokussieren die Interpretation auf Veranstaltungsebene (siehe Manuskript 1 in Abschnitt 4) bzw. berücksichtigen die Mehrfachevaluationen (siehe Manuskript 2 in Abschnitt 5). Dadurch gehen 33 der 35 Items des PELVE-Fragebogens in das Messmodell ein. Alle Analysen dieser Arbeit stützen sich auf das in Abbildung 2.5 dargestellte Messmodell. Tabelle 2.4 zeigt die Zuordnung der Items mit Variablennamen und Itemtext zu den einzelnen Bewertungsdimensionen.

Tabelle 2.4: *Relevante Items des PELVE-Fragebogens für das Messmodell*

Zuordnung der Items zu den 7 Bewertungsdimensionen des PELVE-Fragebogens	
Gesamteindruck (*sal*)	
Item	Itemtext (Antwortskala: 1=*stimme nicht zu*, ... ,5=*stimme zu*)
sal_1	Die Veranstaltung trägt zu meinem Interesse am Thema bei.
sal_2	Der behandelte Stoff knüpft an meinen bisherigen Wissensstand an.
sal_3	Die Veranstaltung versetzt mich in die Lage, die Inhalte selbstständig zu vertiefen.
sal_4	Das fachliche Niveau der Veranstaltung empfinde ich als angemessen.
sal_5	Kommilitonen würde ich den Besuch dieser Veranstaltung empfehlen.
sal_6	Insgesamt gesehen, bin ich mit dieser Lehrveranstaltung zufrieden.
Fachkompetenz (*squA*)	
Item	Itemtext (Antwortskala: 1=*wenig*, ... ,5=*viel*)
Ich habe durch den Besuch dieser Lehrveranstaltung folgende Qualifikationen erworben:	
squ_1	Wissen über Theorien und Modelle
squ_2	Wissen über Fakten, Begriffe und Konzepte
squ_3	Wissen über Forschungsverfahren und wissenschaftliche Methoden
squ_4	Anwendung von Theorien, Methoden, Konzepten
sonstige Komptenzen (*squB*)	
Item	Itemtext (Antwortskala: 1=*wenig*, ... ,5=*viel*)
Ich habe durch den Besuch dieser Lehrveranstaltung folgende Qualifikationen erworben:	
squ_5	Praxiswissen, tätigkeitsrelevantes Wissen
squ_6	Schlüsselkompetenzen (Präsentieren, Arbeiten im Team, Recherchieren, ...)
squ_7	Kompetenz zu unabhängigem und selbstständigem Arbeiten
squ_8	Fachübergreifendes Denken
allgemeine Rahmenbedingungen (*sraA*)	
Item	Itemtext (Antwortskala: 1=*stimme nicht zu*, ... ,5=*stimme zu*)
sra_1	Die räumlichen Gegebenheiten (Größe, bauliche Qualität, Lage, ...) sind für diese Veranstaltung ausreichend.
sra_2	Die Ausstattung (Medien, Technik, Modelle, ...) ist für diese Veranstaltung angemessen.
sra_3	Die Veranstaltung findet in einem angemessenen zeitlichen Rahmen (Zeitpunkt, Dauer, Überschneidungen, ...) statt.
Begleitmaterialien (*sraB*)	
Item	Itemtext (Antwortskala: 1=*stimme nicht zu*, ... ,5=*stimme zu*)
sra_4	Begleitmaterialien (Literatur, Skript, ...) stehen in ausreichendem Maße zur Verfügung.
sra_5	Die verfügbaren Begleitmaterialien (Literatur, Skript, ...) sind hilfreich.

Zuordnung der Items zu den 7 Bewertungsdimensionen des PELVE-Fragebogens. (Fortsetzung)

Dozentenverhalten (sdo)	
Item	Itemtext (Antwortskala: 1=*stimme nicht zu*, ... ,5=*stimme zu*)
Der Dozent/die Dozentin ...	
sdo_1	hat Ziele und Struktur der Veranstaltung nachvollziehbar dargestellt.
sdo_2	geht, soweit möglich, auf organisatorische Wünsche der Teilnehmenden ein.
sdo_3	teilt die Veranstaltungszeit sinnvoll ein (auf Vortrag, Diskussion, Klärung von Fragen, ...).
sdo_4	steht bei Bedarf für Rückfragen und weitere Hilfestellung zur Verfügung.
sdo_5	schafft eine anregende Arbeitsatmosphäre.
sdo_6	bereitet die Einzelsitzungen angemessen vor.
sdo_7	greift inhaltliche Anregungen und Fragen der Teilnehmenden auf.
sdo_8	ordnet Einzelaspekte in einen thematischen Gesamtzusammenhang ein.
sdo_9	Insgesamt gesehen, bin ich mit dem Beitrag des Dozenten/der Dozentin zu dieser Lehrveranstaltung zufrieden.
Studentenverhalten (ste)	
Item	Itemtext (Antwortskala: 1=*stimme nicht zu*, ... ,5=*stimme zu*)
Die meisten Teilnehmenden dieser Lehrveranstaltung ...	
ste_1	besuchen die Veranstaltung regelmäßig.
ste_2	bereiten sich auf die einzelnen Termine angemessen vor.
ste_3	beteiligen sich, soweit möglich, aktiv an der Veranstaltung.
ste_4	verfolgen die Veranstaltung aufmerksam und mit Interesse.
ste_5	Insgesamt gesehen, bin ich mit dem Verhalten der meisten Teilnehmenden zufrieden.

Anmerkungen. Zuordnung der 33 Items zu den Bewertungsdimensionen (Facetten) des PELVE-Fragebogens, die in die Analysen der vorliegenden Arbeit eingehen. In der Spalte *Item* sind die Variablennamen der entsprechenden Items zu finden und in der Spalte *Itemtext* ist der exakte Wortlaut des Items aufgelistet. Der vollständige Fragebogen ist im Anhang A dargestellt.

2.2.3 Ergebnisberichte der LVE

Im Universitätsprojekt Lehrevaluation nehmen die Ergebnisberichte einen besonderen Stellenwert ein. Während die Beschreibung der LVE (vgl. Abschnitt 2.1) den Evaluationsprozess und die Ergebnisse fokussiert, bleibt der Umgang mit den Ergebnissen außen vor. Das Ziel von LVE ist nicht nur die Erhebung des Status Quo. Vielmehr ist ihr Feedbackcharakter darauf ausgelegt, die Lehre zu verbessern. Die Aufbereitung der Ergebnisse in einem Bericht sind daher ein bedeutsamer Bestandteil in der LVE und ein übliches Vorgehen (Spooren, 2012). Auf Basis der Ergebnisberichte kann eine Beurteilung der Lehre erfolgen und Aufschluss darüber geben, ob Lehre ihre intendierten Effekte erreicht (Wolbring, 2013). Können die Ergebnisberichte zur Verbesserung der Lehre eingesetzt werden, wird die LVE als solches stärker akzeptiert, als wenn dies nicht der Fall ist (vgl. Cohen, 1980). Mit Hilfe einer durchdachten Struktur ermöglicht das ULe mit den Ergebnisberichten der LVE diese Ziele

zu erreichen. In Studie 3 wird näher untersucht, ob die Variation einzelner Elemente des Ergebnisberichts zur Verbesserung der Lehre beitragen kann (vgl. Manuskript 3 in Abschnitt 6). Hierfür gibt es bisher kontroverse Ansätze. Auf der einen Seite ist der Ergebnisbericht die einzige und umfassendste Quelle zur Dokumentation der LVQ, auf der anderen Seite werden die damit verbunden Effekte als sehr gering betrachtet (Marsh, 2007a; Rindermann, 2009). Der bloße Empfang der Ergebnisberichte reicht selbstverständlich nicht aus. Dozenten müssen sich mit den Ergebnissen auseinandersetzen. Diese Aspekte werden im Rahmen der Rezeptionsforschung untersucht. Die Art und Weise der Ergebnisaufbereitung liegt in der Verantwortung der Evaluationsbeauftragten und in diesem Fall bei ULe. Es liegt nahe, dass der Rezeptionsprozess durch die Darstellungen der Ergebnisse zusätzlich beeinflusst werden kann. Eine möglichst zeitnahe Rückmeldung der Ergebnisse und eine ansprechende Aufbereitung (grafische Darstellungen, Länge des Berichts, usw.) können dazu beitragen, dass die Ergebnisse vom Dozenten mit einer höheren Wahrscheinlichkeit gelesen werden, als wenn der Bericht sehr lang ist und erst spät dem Dozenten zur Verfügung gestellt wird. In Studie 3 werden die Darstellungen im Bericht variiert. Im Rahmen eines randomisierten Experiments werden zwei unterschiedliche Versionen verglichen, die sich in ihrer Komplexität und Informationsdichte unterscheiden. Die komplexere Version mit ihrer höheren Informationsdichte hat den Vorteil, dass der Bericht kürzer ist als die einfachere Version. Außerdem werden alle relevanten Verteilungsinformationen an einem Ort dargestellt und nicht, wie in der einfachen Version, in getrennten Berichtteilen. Allerdings könnten komplexere Darstellungen auch falsch interpretiert werden und zu Missverständnissen führen. Obwohl der Einfluss verschiedener Ergebnisdarstellungen plausibel erscheint, widmet man sich in der Hochschulforschung vorrangig der Fragebogenkonstruktion und weniger den Aspekten der Rezeption von Evaluationsergebnissen. Änderungen am Ergebnisbericht können unter Umständen negative Effekte auf die Rezeption und damit auf die Interpretation der Ergebnisse haben. Um diese möglicherweise negativen Effekte einer neuen Berichtversion empirisch zu untersuchen, wurde Studie 3 durchgeführt (vgl. Manuskript 3 in Abschnitt 6).

Die Entwicklung einer neuen Berichtversion gründet sich auf drei Vorstudien zur Rezeption von LVE-Ergebnissen (vgl. Vetterlein & Sengewald, 2011). In der ersten Vorstudie wurden standardisierte Interviews über die LVE-Ergebnisse mit den Dozenten geführt. Anschließend erhielten die Dozenten einen Onlinetest, um zu untersuchen, welche Elemente der Grafiken für eine Interpretation der Ergebnisse genutzt werden. Die zweite Vorstudie beinhaltete eine Befragung der Studenten und fokussierte die Ergebnisrückmeldung aus Sicht der Studenten. In der dritten Vorstudie wurden ebenfalls Interviews mit Dozenten geführt und fünf neue Ergebnisgrafiken vergleichend evaluiert. Insgesamt

gaben nur 13 % der befragten Studenten an, dass sie im vergangenen Semester an keiner Evaluation teilgenommen haben. Ein Großteil der Studenten hat mindestens eine Veranstaltung im vorangegangenen Semester evaluiert. Von den Studenten, die an einer Evaluation teilgenommen haben, gaben zudem nur 14 % an, dass sie die Evaluationsergebnisse nicht kennen. In diesen Fällen erfolgte keine Rückmeldung der Ergebnisse von den Dozenten an die Studenten. In den übrigen Fällen wurden die Ergebnisse an die Studenten zurückgemeldet. Dies setzt ein Mindestmaß an Rezeption der Ergebnisse durch den Dozenten voraus. Eine übliche Annahme, dass sich die Dozenten mit ihren Evaluationsergebnissen im Allgemeinen nicht beschäftigen, kann hier nicht bestätigt werden. In 73 % der Fälle wurden die Ergebnisse mit visueller Unterstützung den Studenten präsentiert. Zusätzlich gaben 45 % der Studenten, die eine Ergebnispräsentation erhalten haben an, dass diese zwischen 10 und 20 Minuten in Anspruch nahm. Damit kann von einer zum Teil intensiven Auseinandersetzung mit den Ergebnissen ausgegangen werden. Die Vorstudien deuten allesamt darauf hin, dass an der FSU Jena eine Auseinandersetzung mit den Evaluationsergebnissen stattfindet. Damit ist die wichtigste Voraussetzung für die Rezeption von Ergebnissen gegeben, die Wahrnehmung der Ergebnisse. Unklar ist weiterhin, welche Ergebnisse genau präsentiert wurden. Wenn nur positive Ergebnisse an die Studenten zurückgemeldet werden, ist fraglich, ob kritische Punkte zu einer Veränderung der Lehre führen. Auch ist weiterhin unklar, wie viele Dozenten, bei denen theoretisch Änderungsbedarf bestünde, ihre Lehrveranstaltung evaluieren lassen und die Ergebnisse zurückmelden.

Um zu prüfen, welche Änderungen am Ergebnisbericht gewünscht sind, wurden eine Reihe von Entwürfen in der dritten Vorstudie pilotiert. Die Entwürfe unterschieden sich im Hinblick auf die Darstellung des Mittelwerts und die Varianz der Studentenurteile. In einem Entwurf wurde die Varianz über einen Fehlerbalken eingezeichnet und die Antwortverteilung separat am Ende des Berichts aufgeführt (einfache Darstellung). Ein anderer Entwurf integrierte die Antwortverteilung direkt in die Grafik, in der auch der Mittelwert dargestellt wurde und verzichtete auf den Fehlerbalken (komplexe Darstellung). Die Ergebnisse dieser Vorstudie zeigen die Vorteile beider Grafiktypen auf. Zwar empfanden die Dozenten die einfache Darstellung als gut und leicht verständlich (Vetterlein & Sengewald, 2011), dennoch nutzten die Dozenten verstärkt die Antwortverteilung in der Interpretation der Ergebnisse, wenn sie in derselben Grafik wie der Mittelwert dargeboten wurde (komplexe Darstellung). Die Verwendung der Antwortverteilung bei der Interpretation ist sehr zu empfehlen. Hier wird deutlich, wie viel Prozent der Studenten zustimmen oder den Aspekt des Items ablehnen. Die Verwendung der Antwortverteilung zur Interpretation der Varianz von Studentenurteilen fällt den Dozenten zudem deutlich leichter als die Interpretation auf Basis der Standardabweichung (Vetterlein & Sen-

gewald, 2011). Ob die damit gewählte neue Ergebnisgrafik für die Darstellung der Ergebnisse geeignet ist, kann auf Basis der Vorstudien nicht abschließend geklärt werden. Durch die Komprimierung der Informationen in der neuen Datengrafik sind durchaus auch Missverständnisse möglich, die zu falschen Interpretationen, Entscheidungen und schlechteren Lehrveranstaltungen führen könnten. Aus diesem Grund wurde in Studie 3 (vgl. Manuskript 3 in Kapitel 6) ein Experiment durchgeführt, das die LVE-Ergebnisse zwischen Veranstaltungen mit neuem (komplexe Darstellung) und altem (einfache Darstellung) Bericht vergleicht.

3 Übersicht zu den Manuskripten

Die vorliegende Arbeit beinhaltet drei Manuskripte, die im Folgenden eingebunden sind. Manuskript 1 und 3 sind bereits in der Zeitschrift Diagnostica publiziert. Manuskript 2 wurde zum Zeitpunkt der Einreichung dieser Arbeit zur Publikation in der Zeitschrift Diagnostica eingereicht und befindet sich in Begutachtung. Zu jedem Artikel wird der Eigenanteil der beteiligten Autoren dargestellt.

3.1 Manuskript 1

Sengewald, E. & Vetterlein, A. (2015). Multilevel Faktorenanalyse für Fragebögen zur Lehrveranstaltungsevaluation. *Diagnostica, 61*, 116–123.

3.1.1 Zusammenfassung

Das erste Manuskript umfasst eine Studie zur Untersuchung des Messmodells des Fragebogens PELVE. Die konfirmatorische Prüfung des postulierten Messmodells wird unter Verwendung verschiedener Verfahren zur konfirmatorischen Faktorenanalyse (CFA-Verfahren) durchgeführt. Dabei steht die konfirmatorische Multilevel-Analyse (ML-CFA) im Vordergrund der Arbeit. Sie wird bisher nur selten für die Überprüfung des Messmodells von Fragebögen zur LVE eingesetzt, obwohl sie das theoretische Modell der LVQ und das Erhebungsdesign am besten abbilden kann. In dem Manuskript werden drei CFA-Verfahren eingeführt und am Beispiel des PELVE-Fragebogens wird untersucht, welches Verfahren den besten Modellfit zeigt.

3.1.2 Eigenanteil der Autoren

Die grundlegende Idee zur Überprüfung des Messmodells für den PELVE-Fragebogen entstand in gemeinsamer Diskussion von Erik Sengewald und Anja Vetterlein. Literaturrecherche, die Wahl der CFA-Verfahren zur Prüfung des Messmodells und die Modellherleitung wurden durch Erik Sengewald eigenständig durchgeführt. Die Datenanaufbereitung und -analyse sowie die Manuskripterstellung

wurde ebenfalls durch Erik Sengewald eigenständig durchgeführt. Anja Vetterlein war überwiegend am Review des Manuskripts sowie an theoretischen Diskussionen beteiligt.

3.2 Manuskript 2

Sengewald, E. & Vetterlein, A. (2015). Einfluss der Mehrfachevaluation auf die Ergebnisse konfirmatorischer Faktorenanalysen in der Lehrveranstaltungsevaluation. Manuskript eingereicht zur Publikation am 26.08.2015.

3.2.1 Zusammenfassung

In der Einleitung (siehe Abschnitt 1) wurde das Thema der Mehrfachevaluation bereits eingeführt. In Manuskript 2 werden die Datenstrukturen von LVE-Stichproben genauer erläutert und die Mehrfachevaluation im Kontext von Multiple-Membership-Modellen eingeführt. Das Manuskript geht auf die Konsequenzen der Mehrfachevaluation für die Kennwerte der Modellpassung im Rahmen verschiedener konfirmatorischer Faktorenanalysen ein. Es wird an Hand empirischer Daten gezeigt, dass Mehrfachevaluation bei der konfirmatorischen Faktorenanalyse eine Quelle für eine schlechtere Modellpassung sein kann. Der Einfluss der Mehrfachevaluation auf die Modellfitmaße wird am Beispiel des Messmodells für den Fragebogen PELVE untersucht. Hierfür werden drei CFA-Verfahren eingesetzt, um diese Effekte für unterschiedliche, in der LVE übliche, Verfahren zu untersuchen. Abschließend wird der Einfluss der Mehrfachevaluation auf die Schätzung von Faktorwerten untersucht und die Verwendung der LVE zum Ranking von Veranstaltungen thematisiert.

3.2.2 Eigenanteil der Autoren

Die Datenaufbereitung, Durchführung der konfirmatorischen Faktorenanalysen, Ergebnisauswertung und Theoriearbeit wurden durch Erik Sengewald vollständig und eigenständig durchgeführt. Das Manuskript wurde ebenfalls eigenständig durch Erik Sengewald verfasst. Anja Vetterlein unterstützte die Überarbeitung des Manuskripts und gab hilfreiche Hinweise zur theoretischen Einbettung.

3.3 Manuskript 3

Vetterlein, A. & Sengewald, E. (2015). Ergebnisdarstellung in der Lehrveranstaltungsevaluation. Effekte verschiedener Berichte auf die Qualität von Lehrveranstaltungen. *Diagnostica*, *61*, 153–162.

3.3.1 Zusammenfassung

Manuskript 3 beinhaltet eine Studie zur Untersuchung spezifischer Eigenschaften des Ergebnisberichts. Hierfür wurden die Datengrafiken entsprechend den Erkenntnissen aus Vorstudien (vgl. Seite 29) erstellt und im Rahmen einer Treatmentstudie untersucht. Im Manusript 3 wird diese Treatmentstudie näher erläutert. Aufbauend auf den Ergebnissen zur ML-CFA erfolgt die Auswertung auf Basis der latenten Variablen auf Veranstaltungsebene. Die Ergebnisse verdeutlichen, dass keine negativen Effekte auftreten, obwohl die neue Datengrafik durch eine hohe Informationsdichte deutlich komplexer ist. Die gefundenen positiven Effekte auf die LVE-Ergebnisse der nachfolgenden Lehrveranstaltung des Dozenten werden diskutiert und Empfehlungen für den Umgang mit Weiterentwicklungen von Ergebnisberichten thematisiert.

3.3.2 Eigenanteil der Autoren

Anja Vetterlein und Erik Sengewald haben gemeinsam neue Ergebnisgrafiken entwickelt. Erik Sengewald hat überwiegend die Durchführung der Vorstudien geleitet. Die Auswertung der Vorstudien und die Analysen für das vorliegende Manuskript übernahm überwiegend Anja Vetterlein. Bei der Datenaufbereitung und -analyse wurde sie von Erik Sengewald unterstützt. Die Durchfürhung der Treatmentstudie, die Datenanalyse und die Manuskripterstellung erfolgte eigenständig von Anja Vetterlein. Erik Sengewald war in Teilen an der Überarbeitung des Manuskripts beteiligt.

4 Manuskript 1

Diagnostica, 61 (3), 116 – 123 © Hogrefe Verlag, Göttingen 2015

Multilevel Faktorenanalyse für Fragebögen zur Lehrveranstaltungsevaluation

Erik Sengewald und Anja Vetterlein

Zusammenfassung. Zur Überprüfung der postulierten Messmodelle für Fragebögen zur Lehrveranstaltungsevaluation (LVE) werden konfirmatorische Faktorenanalysen (CFA) entweder auf Studenten- oder Veranstaltungsebene durchgeführt. Der resultierende Modellfit ist oft inakzeptabel (Marsh et al., 2009). Die vorliegende Studie vergleicht die konventionellen CFA-Verfahren in der LVE mit einer Multilevel-CFA anhand eines empirischen Beispiels mit 183 334 Studentenurteilen. Die Ergebnisse verdeutlichen die Überlegenheit der Multilevel-CFA für die Analyse des Messmodells eines Fragebogens zur LVE. Es werden die Vorteile der Multilevel-CFA für die Erfassung der multidimensionalen Veranstaltungsqualität aufgezeigt und Anwendungsmöglichkeiten für die Hochschulforschung diskutiert.
Schlüsselwörter: Hochschulforschung, Multilevel Modelle, konfirmatorische Faktorenanalyse, Veranstaltungsqualität, Lehrevaluation

Multilevel Factor Analysis for Students' Evaluations of Teaching

Abstract. The measurement model of questionnaires for students' evaluations of teaching (SET) is typically evaluated with confirmatory factor analysis (CFA) using either student ratings or class means. However, the modelfit is often inacceptable (Marsh et al., 2009). We compare the traditional CFA approaches of SETs with the multilevel confirmatory factor analysis (ML-CFA) based on an empirical example of 183 334 student ratings. The application of a ML-CFA to SETs is strongly recommended due to the results. We emphasize the advantages of a ML-CFA for measuring course quality and discuss applications for research in higher education.
Keywords: higher education, multilevel models, confirmatory factor analysis, course quality, students' evaluation of teaching

Die Qualität einer Lehrveranstaltung ist ein mehrdimensionales Konstrukt (vgl. Marsh et al., 2009; Rindermann, 2009; Spooren, Brockx & Mortelmans, 2013). Zur Messung der mehrdimensionalen Eigenschaften einer Lehrveranstaltung werden *Fragebögen zur Lehrveranstaltungsevaluation* (LVE) eingesetzt. Der Anspruch an diese Fragebögen ist hoch. Sie müssen möglichst kurz sein, die Multidimensionalität guter Lehre erfassen und den psychometrischen Kriterien guter Messmodelle genügen (vgl. Marsh et al., 2009). Zur Überprüfung der theoretischen Messmodelle eignen sich *konfirmatorische Faktorenanalysen* (CFA). Im Rahmen der LVE-Forschung werden im Wesentlichen drei Ansätze verfolgt: (a) CFA auf Studentenebene (Rohdatenanalyse), (b) CFA auf Veranstaltungsebene (Mittelwertanalyse), (c) CFA auf Studenten- und Veranstaltungsebene *(Multilevel-CFA)*.

Dieser Artikel wurde im Rahmen des Gemeinsamen Bund-Länder-Programms für bessere Studienbedingungen und mehr Qualität in der Lehre aus Mitteln des Bundesministeriums für Bildung und Forschung (BMBF) unter dem Förderkennzeichen 01PL12071 gefördert. Wir danken Prof. Dr. Matthias Ziegler, Humboldt-Universität zu Berlin, für seine Unterstützung und Marie-Ann Sengewald, Friedrich-Schiller-Universität Jena, für wertvolle Anmerkungen zu früheren Versionen dieses Manuskripts.

DOI: 10.1026/0012-1924/a000140

Bisher kommen vorrangig die Verfahren (a) und (b) zum Einsatz (eine Übersicht hierzu findet sich in d'Apollonia & Abrami, 1997; Rindermann, 2009; Spooren et al., 2013). Diese CFAs der LVE-Fragebögen zeigen häufig keine zufriedenstellende Modellpassung, wie Marsh et al. (2009) anführen: „Conventional CFA goodness of fit criteria are too restrictive when applied to most multifactor rating instruments. […] it is almost impossible to get an acceptable fit (e. g., CFI, RNI, TLI $> .9$; RMSEA $< .05$) for even good multifactor rating instruments […]" (S. 441).

In Folge der schlechten Modellpassung werden häufig Items revidiert, aus dem Messmodell entfernt oder die Modellpassung wird explorativ (z. B. mit einer exploratorischen Faktorenanalyse) beurteilt (vgl. Marsh et al., 2009; Rindermann, 2009).

Die Multilevel-CFA, Verfahren (c), ist im Falle der LVE das theoretisch angemessene Verfahren, weil unter Verwendung der Erhebungseinheit (Studentenurteile) latente Variablen auf Veranstaltungsebene (Analyseeinheit) konstruiert werden können. Bisher wird die Multilevel-CFA (ML-CFA) nur selten angewandt (vgl. Toland & de Ayala, 2005). Die vorliegende Studie folgt dem Vorschlag von Rindermann (2009) und vergleicht die drei Ansätze

zur Modellgeltungskontrolle für Messmodelle der LVE. Es wird die Frage beantwortet, welche Konsequenzen bei der Verwendung konventioneller Verfahren zur Modellgeltungskontrolle im Vergleich zur ML-CFA beobachtbar sind. Der Vergleich der CFA-Verfahren erfolgt anhand des Fragebogens PELVE (*Prozess- und Ergebnisorientierte Lehrveranstaltungsevaluation*; vgl. Born, Loßnitzer & Schmidt, 2006; Loßnitzer, Schmidt & Born, 2007).

Theorie

Fragebögen zur LVE werden auf Grundlage verschiedener Theorien zur Lehrqualität bzw. Veranstaltungsqualität konzipiert. Je nach Instrument werden im deutschsprachigen Raum 21 bis 66 Items zur Messung der 3 bis 21 postulierten Faktoren der Lehr- und Veranstaltungsqualität erfasst (vgl. Rindermann, 2009). In Schmidt und Loßnitzer (2010) findet sich eine ausführliche Darstellung, welche Konstrukte von verschiedenen Fragebögen zur LVE erfasst werden. Der Fragebogen PELVE (Born et al., 2006; Loßnitzer et al., 2007) erfüllt die Kriterien der Definition von Lehrveranstaltungsevaluation nach Schmidt und Loßnitzer (2010) und vereint Prozess- und Ergebnisvariablen in einem Fragebogen.

PELVE. Der Fragebogen PELVE wurde im Universitätsprojekt Lehrevaluation der Friedrich-Schiller-Universität Jena (FSU Jena) entwickelt und wird seit 2004 für die LVE eingesetzt. Insgesamt umfasst der Fragebogen 35 veranstaltungsübergreifende Ratingitems, die durch spezifische Items für Vorlesungen, Seminare und Übungen ergänzt werden. Demographische Angaben, Items zu Arbeitsaufwand, freitextliche Anmerkungen und optionale Items vervollständigen den Fragebogen. Die Ratingitems lassen sich sieben Facetten guter Lehre zuordnen, wobei vier (*Dozentenverhalten*, *Studentenverhalten*, *Rahmenbedingungen* und *Begleitmaterialien*) zur Evaluation des Lehrprozesses und drei (*Gesamteindruck*, *Fachkompetenz* und *sonstige Kompetenzen*) zur Evaluation des Lehrergebnisses herangezogen werden (vgl. Loßnitzer et al., 2007). Eine ausführliche Beschreibung der Dimensionen des PELVE befindet sich auch in Born et al. (2006).

Items der Kompetenzdimensionen werden auf einer fünfstufigen Likert-Skala mit den Antwortpolen *wenig* bis *viel* beantwortet. Alle anderen Ratingitems sind ebenfalls fünfstufig mit den Antwortpolen *stimme nicht zu* bis *stimme zu.*

Mehrebenenstruktur per Design. Bei einer LVE sollen Aussagen über die Qualität einer Lehrveranstaltung durch den Einsatz einer standardisierten Befragung unter den Studenten getroffen werden (Schmidt & Loßnitzer, 2010). Das Design jeder LVE lässt sich demnach als Fremdeinschätzung der Veranstaltung durch ihre Studenten darstellen. Durch die Gruppierung der Studenten in Veranstaltungen ist die Mehrebenenstruktur per Design gegeben. Sofern die Eigenschaften der Veranstaltung einen Einfluss auf die Ratings der Studenten haben, werden Ratings derselben Veranstaltung größere Ähnlichkeit aufweisen als Ratings verschiedener Veranstaltungen. Dennoch können sich die Ratings innerhalb einer Veranstaltung unterscheiden. Im Rahmen der LVE wird daher angenommen, dass Eigenschaften der Veranstaltung und der Student als Rater das Ergebnis der LVE beeinflussen. Wenn diese Annahme korrekt ist, ist die Berücksichtigung der Mehrebenenstruktur auch bei Tests der Varianz-Kovarianz-Matrix bzw. der Überprüfung des Messmodells nötig.

Konfirmatorische Faktorenanalyse (CFA). Die CFA kann auf Grundlage der individuellen Studentenurteile erfolgen (CFA auf Studentenebene) oder nach Aggregation der Studentenurteile auf Grundlage von Veranstaltungsmittelwerten (CFA auf Veranstaltungsebene; vgl. Rindermann, 2009). Im Gegensatz dazu berücksichtigt die ML-CFA die Studenten- und Veranstaltungsebene (vgl. Toland & de Ayala, 2005).

CFA auf Studentenebene. Bei der CFA auf Studentenebene werden die Rohdaten zur Analyse des Messmodells verwendet. Die Gruppierung der Studenten in Veranstaltungen wird dabei vernachlässigt. Die so konstruierten latenten Variablen sind damit auf Personenebene und nicht auf Veranstaltungsebene definiert. Diese CFA basiert auf der Annahme identisch verteilter und unabhängiger Beobachtungen (*iid-Annahme*; vgl. hierzu Muthén & Satorra, 1995; Skinner, Holt & Smith, 1989). Die Richtigkeit dieser Annahme im Rahmen der LVE wird von Toland und de Ayala (2005) angezweifelt und bedarf einer empirischen Prüfung. Die *Intraklassenkorrelation* (ICC) verdeutlicht das Ausmaß der Verletzung dieser Annahme und misst die Ähnlichkeit der Studierendenurteile innerhalb derselben Veranstaltung (Muthén & Satorra, 1995). Julian (2001) zeigt in seiner Simulationsstudie, dass bei ICC-Werten ab .15 die Modellparameter der CFA überschätzt, die Standardfehler unterschätzt und die χ^2-Werte zu hoch sind, wenn die hierarchische Struktur der Daten nicht berücksichtigt wird.

CFA auf Veranstaltungsebene. Um den Einfluss hoher ICCs auf die Modellpassung zu umgehen, kann die CFA auf Veranstaltungsebene durchgeführt werden. Zudem empfehlen viele Autoren (vgl. z. B. Clayson, 2007; Marsh, 1983; Marsh & Roche, 1997; Rindermann, 2009) diesen Ansatz der CFA, um interpretierbare Aussagen auf Veranstaltungsebene zu erhalten. Folglich basiert diese CFA auf den Mittelwerten der Items innerhalb der Veranstaltung. Die individuellen Urteile der Studierenden gehen nicht in das Messmodell ein. Die vorliegende Varianz der Ratings innerhalb der Veranstaltung wird vernachlässigt. Dadurch werden jedoch Zusammenhänge auf der Veranstaltungsebene überschätzt (vgl. Kaplan & Elliot, 1997).

Multilevel-CFA. Mit Hilfe der ML-CFA ist es möglich unter Verwendung der Erhebungseinheit (Studentenebene) Aussagen über latente Variablen auf der interes-

sierenden Analyseebene (Veranstaltungsebene) zu treffen. Dafür wird ein Messmodell auf Studentenebene (Within-Messmodell) und auf Veranstaltungsebene (Between-Messmodell) erstellt (vgl. Asparouhov & Muthén, 2006; Muthén, 1994; Muthén & Asparouhov, 2015; Toland & de Ayala, 2005).

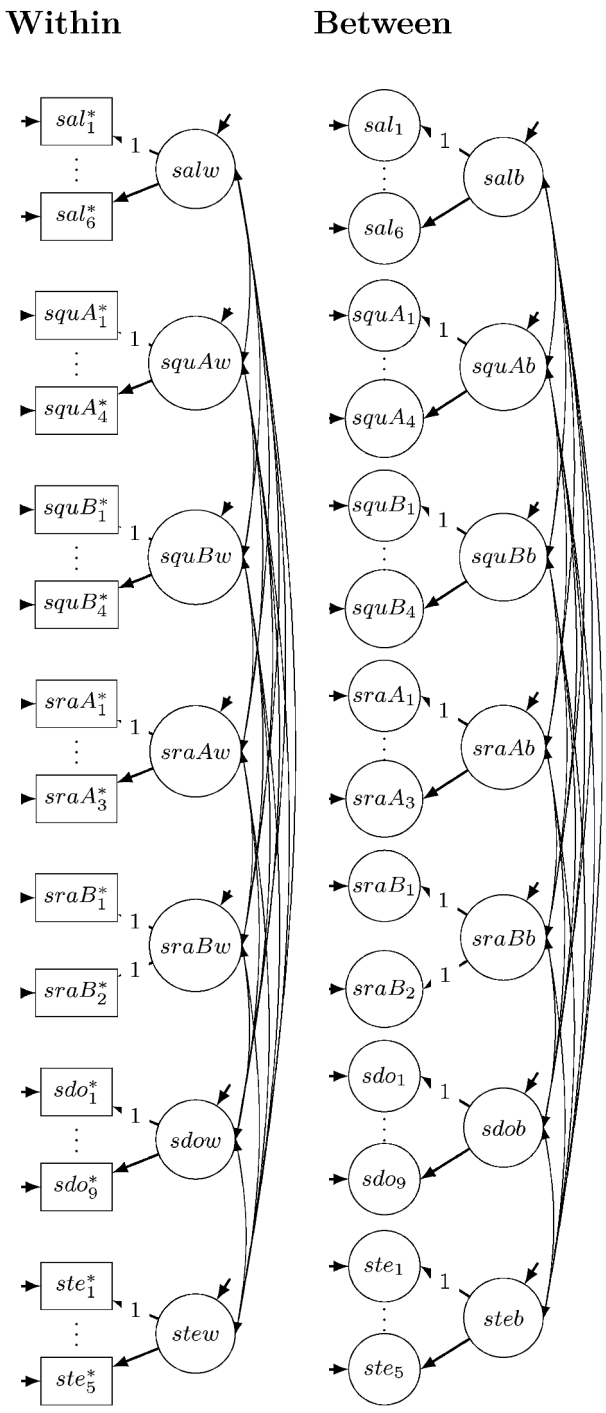
Die ML-CFA ist aus theoretischer Sicht das beste Verfahren zur Überprüfung des Messmodells, weil es das Design der LVE genauer abbildet als die anderen Verfahren. Inwiefern die Verwendung der konventionellen CFA-Verfahren entweder auf Studentenebene oder auf Veranstaltungsebene Konsequenzen für die Modellpassung hat, wird am Beispiel des Messmodells für den Fragebogen PELVE und den Daten zur LVE der FSU Jena überprüft.

Methode

In der vorliegenden Studie werden die verschiedenen Ansätze zur CFA am Beispiel des PELVE miteinander verglichen. In Abbildung 1 ist das Multilevel-Messmodell des PELVE schematisch dargestellt. Es werden je sieben latente Variablen auf Studenten- und Veranstaltungsebene postuliert. Die Single-Level-Messmodelle (CFA auf Studenten- bzw. Veranstaltungsebene) entsprechen grafisch dem Within-Modell aus Abbildung 1. Für die Prüfung des Messmodells werden die 33 Ratingitems des PELVE verwendet, die nach Born et al. (2006) theoriegeleitet jeweils genau einer Dimension zugeordnet werden können und in Vorlesungen, Seminaren und Übungen zum Einsatz kommen.

Stichprobe. Es gehen alle LVE der FSU Jena vom Sommersemester 2005 bis 2013 in die Analyse ein. Insgesamt sind $N_V = 7\,459$ Veranstaltungen mit $N_D = 1\,603$ unterschiedlichen Dozenten und $N_S = 183\,334$ Studenten in der Gesamtstichprobe enthalten. Die Dozenten (41 % weiblich) sind im Durchschnitt 36 Jahre ($SD = 10.4$) und die Studenten (64 % weiblich) im Durchschnitt 22.4 Jahre ($SD = 10.5$) alt. In jeder LVE sind mindestens 5 und durchschnittlich $N_{Sv} = 24.58$ Studenten enthalten. Die Gesamtstichprobe wird zusätzlich in eine Vorlesungs-, Seminar- und Übungsstichprobe unterteilt, die sich durch die jeweilige Veranstaltungsform ergibt (vgl. Tabelle 1). In den Stichproben ist die Anzahl Dozenten (N_D) geringer als die Anzahl Veranstaltungen (N_V). Dies ergibt sich aus der Evaluation mehrerer Veranstaltungen je Dozent. Um den Effekt dieser Mehrfachevaluation auf den Modellfit zu schätzen, wird zusätzlich eine Minimalstichprobe betrachtet, in der von $N_D = 100$ zufällig gezogenen Dozenten nur je eine Veranstaltung in die Stichprobe eingeht. Die Beschränkung auf $N_V = 100$ Veranstaltungen stellt eine Untergrenze für eine zuverlässige Parameterschätzung im Rahmen der ML-CFA dar (vgl. Muthén, 1994).

Konfirmatorische Faktorenanalysen. Die konfirmatorischen Faktorenanalysen erfolgen mit der Software M*plus*



Anmerkungen: Gesamteindruck (*sal*), Fachkompetenz (*squA*), sonstige Kompetenzen (*squB*), Rahmenbedingungen (*sraA*), Begleitmaterialien (*sraB*), Dozentenverhalten (*sdo*) und Studentenverhalten (*ste*); Diese Abkürzungen in Kombination mit *w* bzw. *b* verdeutlichen die Modellebene (within bzw. between). Die nummerischen Subskripts repräsentieren die Itemnummer und das Symbol * die latent response-Variable.

Abbildung 1. Schematische Darstellung des Multilevel-Messmodells des Fragebogens zur Prozess- und Ergebnisorientierten Lehrveranstaltungsevaluation (PELVE) mit Studenten- (Within) und Veranstaltungsebene (Between).

Tabelle 1. Stichprobenzusammensetzung

Stichprobe	N_V	N_D	N_S
Gesamt	7 459	1 603	183 334
Vorlesung	1 770	526	74 539
Seminar	3 926	1 043	77 239
Übung	1 763	591	31 556
Minimal	100	100	2 276

Anmerkungen: Dargestellt sind für jede Stichprobe: die Anzahl Veranstaltungen (N_V), die Anzahl Dozenten (N_D) und die Anzahl Studenten (N_S).

7.2 (Muthén & Muthén, 1998–2012). Für jedes Item liegt eine Einfachladung vor, Nebenladungen werden nicht zugelassen (vgl. Abbildung 1). Die Identifizierbarkeit der latenten Variablen wird durch Fixierung der Ladung des jeweils ersten Items auf 1 realisiert. Im Folgenden werden die entsprechenden Verfahren und Messmodelle formal dargestellt. Die Analysen werden für die Gesamtstichprobe, jeden Veranstaltungstyp und die Minimalstichprobe getrennt durchgeführt. Zusätzlich wird eine Multigroup-CFA gerechnet in der die strukturelle und faktorielle Gleichheit des Messmodells für Vorlesungen, Seminare und Übungen angenommen wird.

CFA auf Studentenebene. Aufgrund fehlender multivariater Normalverteilung wird zur Prüfung des Messmodells auf Studentenebene eine CFA für ordinale Variablen unter Verwendung des WLSMV-Schätzers (*weighted least squares mean- and variance-adjusted*; vgl. Muthen, du Toit & Spisic, 1997) durchgeführt. Nach Muthén und Asparouhov (2015) werden zunächst I kontinuierliche latent response Variablen (LRV) Y^*_{pi} ($i = 1, 2, \ldots, I$) angenommen, die einem linearen Messmodell mit D latenten Variablen ϑ_{pd} folgen (vgl. Gleichung 1). Der Index p signalisiert, dass es sich um Variablen auf Personenebene (hier Studentenebene) handelt. Der Zusammenhang zwischen manifester Variable Y_{pi} und LRV Y^*_{pi} wird durch ein Schwellenmodell beschrieben (vgl. Muthén, 1984; Muthén & Asparouhov, 2015). Für Items mit $A = 5$ Kategorien ist das allgemeine Schwellenmodell mit $A - 1$ Schwellen in Gleichung 2 dargestellt. Die manifeste Variable Y_{pi} nimmt einen ihrer A-Werte an, wenn Y_{pi} eine bestimmte Schwelle τ_{ia} ($a = 1, \ldots, A-1$) über- bzw. unterschreitet (vgl. Gleichung 2).

$$Y^*_{p\,i} = \nu_i + \sum_{d=1}^{D} \lambda_{id} \cdot \vartheta_{p\,d} + \varepsilon_{p\,i} \tag{1}$$

$$Y_{p\,i} = \left\{ \begin{array}{lll} 0 & \text{wenn} & Y^*_{p\,i} \le \tau_{i1} \\ 1 & \text{wenn} & \tau_{i1} < Y^*_{p\,i} \le \tau_{i2} \\ 2 & \text{wenn} & \tau_{i2} < Y^*_{p\,i} \le \tau_{i3} \\ 3 & \text{wenn} & \tau_{i3} < Y^*_{p\,i} \le \tau_{i4} \\ 4 & \text{wenn} & \tau_{i4} < Y^*_{p\,i} \end{array} \right\} \tag{2}$$

Gleichung 1 verdeutlicht, dass die Gruppierung der Studenten in Veranstaltungen vernachlässigt wird. Die latenten Variablen ϑ_{pd} sind auf Personenebene und nicht auf Veranstaltungsebene definiert (vgl. Gleichung 1).

CFA auf Veranstaltungsebene. Die CFA auf Veranstaltungsebene beruht auf den Mittelwerten der I Items in J Veranstaltungen. Bezüglich der Variablen Y_{ij} wird ein lineares Messmodell mit D latenten Variablen ϑ_{dj} aufgestellt (vgl. Gleichung 3). Der Veranstaltungsmittelwert $\bar{Y}_{ij}$ eines Items i in der Veranstaltung j wird über den Mittelwert aller Werte der Personen in der Veranstaltung (N_j) errechnet (vgl. Gleichung 4). Die individuellen Urteile der Studenten p in Veranstaltung j auf einem Item i (Y_{pij}) werden im Messmodell selbst nicht berücksichtigt (vgl. Gleichung 3).

$$\bar{Y}_{ij} = \nu_i + \sum_{d=1}^{D} \lambda_{id} \cdot \vartheta_{dj} + \varepsilon_{ij} \tag{3}$$

mit

$$\bar{Y}_{ij} = \frac{1}{N_j} \sum_{p=1}^{N_j} Y_{p\,ij} \tag{4}$$

Multilevel-CFA. Basierend auf dem Schwellenmodell (Gleichung 2) wird ein Messmodell für die latenten Variablen auf Within-Ebene (ϑ_{Wpd}) aufgestellt (vgl. Gleichung 5). Der Index W signalisiert, dass es sich um levelspezifische latente Variablen (ϑ_{Wpd}), Ladungen (λ_{Wid}) und Residuen (ε_{Wpij}) handelt. Das Intercept (ν_{ij}) ist von der Klassenvariable J abhängig. Bezüglich dieses Intercepts wird das Between-Messmodell aufgestellt (vgl. Gleichung 6). In diesem werden latente Variablen auf Veranstaltungsebene (ϑ_{Bdj}) definiert. Das Intercept ν_{ij} aus Gleichung 5 variiert demnach in Abhängigkeit des Wertes der latenten Variablen ϑ_{Bdj} der Veranstaltung $J = j$ auf der Dimension $D = d$. Das explizit formulierte Within- und Between-Messmodell wird im Rahmen der Modellgeltungskontrolle der ML-CFA geprüft.

$$Y^*_{pij} = \nu_{ij} + \sum_{d=1}^{D_W} \lambda_{Wid} \cdot \vartheta_{Wpd} + \varepsilon_{Wpij} \tag{5}$$

$$\nu_{ij} = \nu_i + \sum_{d=1}^{D_B} \lambda_{Bid} \cdot \vartheta_{Bdj} + \varepsilon_{Bij} \tag{6}$$

Beurteilungskriterien. Die Ähnlichkeit der Studierendenurteile innerhalb derselben Veranstaltung wird mithilfe der ICCs beurteilt. Die Berechnung der ICCs erfolgt modellbasiert und ist Teil des Outputs der ML-CFA. Werte über .15 stellen eine substantielle Verletzung der iid-Annahme dar (vgl. Julian, 2001).

Zur Beurteilung der Modellgüte wird der χ^2-Test betrachtet. Bei einem nicht-signifikanten Ergebnis ($p > .05$)

muss das Modell nicht verworfen werden. Der Modellfit kann als akzeptabel betrachtet werden, wenn das Verhältnis aus χ^2-Wert und Freiheitsgraden (χ^2/df) kleiner 2 ist (Schreiber, Stage, King, Nora & Barlow, 2006). Aufgrund der Abhängigkeit des χ^2-Wertes von der Stichprobengröße ist die Minimalstichprobe auf $N_V = 100$ Veranstaltungen beschränkt und es werden die deskriptiven Gütemaße RMSEA, CFI und SRMR zur Beurteilung des Modellfits herangezogen (vgl. Browne & Cudeck, 1993; Hu & Bentler, 1999; Schermelleh-Engel, Moosbrugger & Müller, 2003). Der Modellfit wird als gut interpretiert, wenn der RMSEA $\leq .05$ ist, als akzeptabel für die Werte $.05 <$ RMSEA $\leq .08$, als mittelmäßig für Werte $.08 <$ RMSEA $\leq .10$ und als inakzeptabel für die Werte $.10 <$ RMSEA (vgl. Browne & Cudeck, 1993). Für den CFI zeigen Werte über .97 einen guten Modellfit an, Werte über .95 werden als akzeptabel betrachtet (Schermelleh-Engel et al., 2003). Der SRMR kennzeichnet einen akzeptablen bzw. guten Modellfit bei Werten kleiner als .10 bzw. kleiner als .05. (Hu & Bentler, 1999). RMSEA, CFI und χ^2-Wert sind in *Mplus* 7.2 sowohl für die Single-Level Analysen, als auch für die Multilevel-Analysen verfügbar. Für die ML-CFA stehen zudem der $SRMR_{within}$ und $SRMR_{between}$ zur Verfügung, die eine getrennte Beurteilung des Within- und Between-Messmodells zulassen. Die Simulationsstudie von Ryu und West (2009) zeigt, dass sowohl RMSEA als auch CFI ungeeignet sind, um Missspezifikation des Between-Modells zu identifizieren. Die Beurteilung der Passung des Between-Modells erfolgt daher hauptsächlich auf Basis des $SRMR_{between}$ Index. Für die Bewertung der Modellpassung auf der Within-Ebene eignen sich RMSEA und CFI, da sie adäquat auf eine Missspezifikation des Modells reagieren (vgl. Ryu & West, 2009).

Inwiefern eine unterschiedlich gute Modellpassung für die konkrete Lehrveranstaltung von Bedeutung ist, wird anhand der Faktorwerte untersucht. Hierfür werden die Korrelationen der Faktorwerte berechnet, die im Rahmen der verschiedenen CFA-Verfahren geschätzt werden. Berichtet werden jeweils Mittelwert und Standardabweichung der Korrelationen der Faktorwerte über alle Faktoren des Messmodells. Eine Bewertung der Veranstaltungsqualität erfolgt in der Praxis häufig auf Basis von Rankings. Hierfür werden die geschätzten Faktorwerte der latenten Variablen jeweils in eine Rangreihe transformiert, sodass jede Veranstaltung einen Rangplatz bezüglich einer latenten Variablen hat. In Abhängigkeit des verwendeten CFA-Verfahrens können sich die Rangplätze der konkreten Veranstaltung bzgl. der konkreten latenten Variablen unterscheiden. Berichtet werden die Quartile (*Q1*, *Q2*, *Q3*) der Verteilung der Rangplatzdifferenzen sowie Minimum (*Min.*) und Maximum (*Max.*) dieser Verteilung. Diese Analysen der Korrelationen und Rangplatzdifferenzen werden exemplarisch für die Minimalstichprobe durchgeführt.

Ergebnisse

Die Ergebnisse der CFA auf Veranstaltungs- und Studentenebene sind in Tabelle 2 dargestellt. Das χ^2/df-Verhältnis erreicht bei der CFA auf Veranstaltungsebene keinen akzeptablen Wert. Auch RMSEA und CFI zeigen hier keinen akzeptablen Modellfit an. Folglich muss das postulierte Messmodell nach Prüfung mit der CFA auf Veranstaltungsebene verworfen werden. Die CFA auf Studentenebene lässt keinen eindeutigen Schluss zu. Das χ^2/df-Verhältnis erreicht kein akzeptables Niveau. Im Gegensatz dazu zeigen RMSEA und CFI einen akzeptablen Modellfit an (vgl. Tabelle 2).

Die ML-CFA zeigt sehr gute Ergebnisse (vgl. Tabelle 3). Der RMSEA und die SRMR-Werte beschreiben eine akzeptable Passung des Multilevel-Messmodells für alle Stichproben (vgl. Tabelle 3) und die Multigroup-CFA. Der CFI liegt zum Teil unter der Grenze von .95 für eine akzeptable Passung. Betrachtet man die Minimalstichprobe, kann das χ^2/df-Verhältnis und der CFI = .99 als gut bewertet werden. Der $SRMR_{within}$ zeigt für alle Stichproben eine gute Modellpassung, während der $SRMR_{between}$ als akzeptabel zu interpretieren ist.

Die ICCs der 33 Items variieren zwischen .16 und .44 ($M = .27$, $SD = .06$). Alle ICC-Werte sind damit als hoch zu interpretieren und verweisen auf die Notwendigkeit der ML-CFA und die Gefahr verfälschter Schätzungen von Modellparametern bei Nicht-Berücksichtigung der hierarchischen Datenstruktur (vgl. Julian, 2001). Die Ergebnisse zur Modellpassung der ML-CFA unterstreichen zudem die Überlegenheit der ML-CFA für die Modellgeltungskontrolle im Rahmen der LVE. Dabei zeigt sich unabhängig vom Veranstaltungstyp eine gute Passung des theoretischen Messmodells (Multigroup-CFA). Die konventionellen Ansätze der CFA führen zu unterschiedlichen Beurteilungen bzgl. der Modellpassung. Nach der CFA auf Veranstaltungsebene würde man das Messmodell verwerfen. Bei der CFA auf Studentenebene zeigt sich keine konsistente Richtung der Fitmaße.

Die Korrelationen der Faktorwerte, die auf Basis der unterschiedlichen CFA-Verfahren geschätzt werden, sind hoch aber nicht perfekt. Die Faktorwerte aus der CFA auf Veranstaltungsebene und der ML-CFA korrelieren im Mittel mit .94 ($SD = .03$). Faktorwerte auf Basis der CFA auf Studentenebene korrelieren im Mittel zu .95 ($SD = .03$) mit Faktorwerten auf Basis der ML-CFA und zu .94 ($SD = .06$) mit Faktorwerten auf Basis der CFA auf Veranstaltungsebene. Vergleicht man die Rangplätze nach der Schätzung auf Basis der ML-CFA mit der Schätzung auf Basis der CFA auf Veranstaltungsebene zeigt sich eine breite Verteilung der Rangplatzdifferenzen (*Min.* = -29.7, *Q1* = -4.0, *Q2* = 0.1, *Q3* = 4.6, *Max.* = 27.6). Der Rangplatzvergleich zwischen der CFA auf Studentenebene und der ML-CFA (*Min.* = -27.9, *Q1* = -4.2, *Q2* = 0.2, *Q3* = 4.8, *Max.* = 22.1) bzw. der CFA auf

Tabelle 2. Kennwerte der Modellgüte für die CFA auf Veranstaltungsebene und Studentenebene

Version	Veranstaltungsebene					Studentenebene				
	χ^2	*df*	*p*-Wert	RMSEA	CFI	χ^2	*df*	*p*-Wert	RMSEA	CFI
Gesamtstichprobe	37 044.47	475	.00	.10	.85	332 155.74	475	.00	.06	.95
Vorlesung	11 580.48	475	.00	.12	.83	154 207.08	475	.00	.07	.95
Seminar	18 448.08	475	.00	.10	.85	123 477.59	475	.00	.06	.95
Übung	8 840.13	475	.00	.10	.86	48 000.00	475	.00	.06	.96
Multigruppenanalyse	47 791.96	1 527	.00	.11	.81	386 837.35	1 659	.00	.06	.94
Minimalstichprobe	1 059.63	475	.00	.11	.84	3 675.13	475	.00	.05	.96

Anmerkungen: CFA = confirmatory factor analysis; *df* = Freiheitsgrade; RMSEA = rootmean square error of approximation; CFI = comparative fit index.

Tabelle 3. Kennwerte der Modellgüte für die ML-CFA

Version	χ^2	*df*	*p*-Wert	RMSEA	CFI	$SRMR_{within}$	$SRMR_{between}$
Gesamtstichprobe	281 769.79	953	.00	.04	.90	.04	.07
Vorlesung	95 851.01	953	.00	.04	.92	.05	.07
Seminar	111 020.95	953	.00	.04	.90	.04	.07
Übung	21 662.83	953	.00	.03	.95	.04	.08
Multigruppenanalyse[a]	6 214.56	1 659	.00	.03	.95		
Minimalstichprobe	1 035.95	953	.03	.01	.99	.04	.08

Anmerkungen: df = Freiheitsgrade; ML-CFA = Multilevel confirmatory factor analysis; RMSEA = root mean square error of approximation; CFI = comparative fit index; SRMR = standardized root mean residual. [a] Aufgrund der Limitationen in Mplus 7.2 ist eine Multilevel-Multigroup Analyse unter Verwendung kategorialer Variablen und dem WLSMV-Schätzer nicht möglich. Daher wird hier eine Single-Level CFA durchgeführt und mit TYPE = COMPLEX die Standardfehler und Teststatistik angepasst (vgl. Muthén & Satorra, 1995; Wu & Kwok, 2012).

Veranstaltungsebene (*Min.* = -23.4, *Q1* = -5.0, *Q2* = 0, *Q3* = 4.6, *Max.* = 27.7) zeigt eine ähnliche Verteilung der Rangplatzdifferenzen. Somit sind vereinzelt deutliche Unterschiede im Rangplatz der Veranstaltung je nach verwendeten CFA-Verfahren zu beobachten.

Diskussion

Die Studie vergleicht zwei konventionelle CFA mit der ML-CFA am Beispiel des Fragebogens PELVE. Der Fokus liegt auf der Auswahl der richtigen Methode zur Analyse von Fragebögen zur LVE, weniger auf der Optimierung des Messmodells. In der Überprüfung des theoretischen Messmodells liefert nur die ML-CFA eine adäquate Modellpassung. Dies belegt die Notwendigkeit des Verfahrens, um die komplexe Struktur von LVE-Daten abzubilden und Aussagen auf Veranstaltungsebene zu treffen. Für den PELVE konnte zudem die Generalisierbarkeit des Messmodells auf alle Veranstaltungsformen (Vorlesung, Seminar, Übung) gezeigt werden. Es ist anzunehmen, dass die Ergebnisse auch auf andere Fragebögen zur LVE übertragbar sind, sofern mittlere bis hohe ICCs zu erwarten sind.

Der Vorteil der ML-CFA besteht darin, unter Verwendung der Studentenurteile Aussagen über die latenten Variablen auf Veranstaltungsebene treffen zu können. Damit können u. a. Fragestellungen zur Entwicklung von Veranstaltungsqualität und Wirkung hochschuldidaktischer Maßnahmen untersucht werden. Diese Maßnahmen zielen auf die Veränderung bestimmter Facetten der multidimensionalen Lehr- und Veranstaltungsqualität ab. Während bisherige methodische Ansätze keine adäquate Modellpassung zeigen (CFA auf Veranstaltungsebene) oder keine latenten Variablen auf Veranstaltungsebene definiert sind (CFA auf Studentenebene), kann mittels ML-CFA die Wirkung einer Maßnahme auf die Veranstaltungsqualität eines Dozenten untersucht werden, obwohl Prä- und Postmessung durch unterschiedliche Studenten erfolgen.

Anhand der exemplarischen Untersuchung der Faktorwerte wird deutlich, dass die Verfahren unterschiedliche Ergebnisse im diagnostischen Kontext liefern. So ändert sich im Einzelfall der Rang eines Dozenten sehr stark, wenn der Faktorwert zum Beispiel im Rahmen der ML-CFA anstatt der CFA auf Veranstaltungsebene ermittelt wird. Obwohl die Korrelationen der Faktorwerte zwischen den unterschiedlichen Verfahren auf einem sehr hohen Niveau liegen, sollten die Verfahren nicht vorschnell als diagnostisch gleichwertig betrachtet werden. Gleichzeitig ist bei der Interpretation der Rangplatzdifferenzen darauf zu achten, dass bereits kleine Änderungen der absoluten Faktorwertschätzungen zu größeren Rangplatzdifferenzen führen können. Weitere Studien sollten Aufschluss darüber geben für welche Dimensionen und

unter welchen Bedingungen diese Korrelationen hoch oder niedrig sind. Hierfür sind ausführlichere Untersuchungen notwendig, die auch die Genauigkeit der Punktschätzer für Faktorwerte einbeziehen. In Simulationsstudien kann evaluiert werden inwiefern die Schätzung der Veranstaltungsqualität von der wahren Veranstaltungsqualität in Abhängigkeit des gewählten Verfahrens abweicht.

Grundsätzlich ist über methodische Probleme der ML-CFA im Rahmen der LVE ist noch wenig bekannt. So bedarf es beispielsweise Studien zur Untersuchung möglicher Effekte wiederholter Evaluationen durch dieselben Studenten oder Dozenten, da hierdurch unberücksichtigte Zusammenhänge im Datensatz vorliegen können. Durch die Verwendung einer Minimalstichprobe ohne Mehrfachevaluation liefert die vorliegende Studie erste Hinweise auf den Einfluss dieser Mehrfachevaluation auf den Modellfit. So liegt für die Minimalstichprobe der ML-CFA ein besserer CFI vor als für die Gesamtstichprobe.

Mit Hilfe der ML-CFA liegt ein methodischer Zugang zur Untersuchung der psychometrischen Qualität von Fragebögen zu LVE vor, der den konventionellen Methoden überlegen ist. Fragen zur Reliabilität und Validität dieser Instrumente können damit neu gestellt und die kritische Haltung gegenüber deren Nützlichkeit neu beurteilt werden.

Literatur

Asparouhov, T. & Muthén, B. O. (2006). Comparison of estimation methods for complex survey data analysis. *M*plus *Web Notes.* Zugriff am 11. 12. 2013. Verfügbar unter https://www.statmodel.com/download/SurveyComp21.pdf

Born, S., Loßnitzer, T. & Schmidt, B. (2006). Lehrveranstaltungsevaluation an der Friedrich-Schiller-Universität Jena – Eine Analyse der Dimensionalität der eingesetzten Fragebögen. In B. Krause & P. Metzler (Hrsg.), *Empirische Evaluationsmethoden* (Bd. 10, S. 99 – 116). Berlin: ZeE Verlag.

Browne, M. W. & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136 – 162). Thousand Oaks, CA: Sage.

Clayson, D. E. (2007). Conceptual and statistical problems of using between-class data in educational research. *Journal of Marketing Education*, *29*, 34 – 38.

d'Apollonia, S. & Abrami, P. C. (1997). Navigating student ratings of instruction. *American psychologist*, *52*, 1198 – 1208.

Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1 – 55.

Julian, M. W. (2001). The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling. *Structural Equation Modeling*, *8*, 325 – 352.

Kaplan, D. & Elliot, P. R. (1997). A didactic example of multilevel structural equation modeling applicable to the study of organizations. *Structural Equation Modeling*, *4*, 1 – 24.

Loßnitzer, T., Schmidt, B. & Born, S. (2007). Zentrale Lehrveranstaltungsevaluation an der Friedrich-Schiller-Universität Jena – Qualitätsmodell und Messinstrument. In M. Krämer, S. Preiser & K. Brusdeylins (Hrsg.), *Psychologiedidaktik und Evaluation VI.* (S. 327 – 335). Göttingen: V&R.

Marsh, H. W. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/ course/ instructor characteristics. *Journal of Educational Psychology*, *75*, 150 – 166.

Marsh, H. W., Muthén, B. O., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S. et al. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling*, *16*, 439 – 476.

Marsh, H. W. & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist*, *53*, 1187 – 1197.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika*, *49*, 115 – 132.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*, 376 – 398.

Muthén, B. O. & Asparouhov, T. (2015). Item response modeling in M*plus*: A multi-dimensional, multi-level, and multitimepoint example. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of item response theory: Models, statistical tools, and applications.* Boca Raton, FL: Chapman & Hall/CRC Press. Zugriff am 11. 12. 2013. Verfügbar unter http://www.statmodel.com/download/IRT1Version2.pdf

Muthén, B., du Toit, S. H. C. & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Conditionally accepted for publication in *Psychometrika.*

Muthén, B. O. & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, *25*, 267 – 316.

Muthén, L. K. & Muthén, B. O. (1998 – 2012). *Mplus user's guide* (7th ed.). Los Angeles, CA.

Rindermann, H. (2009). *Lehrevaluation: Einführung und Überblick zu Forschung und Praxis der Lehrveranstaltungsevaluation an Hochschulen mit einem Beitrag zur Evaluation computerbasierten Unterrichts* (2. Aufl.). Landau: Empirische Pädagogik e. V.

Ryu, E. & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling*, *16*, 583 – 601.

Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*, 23 – 74.

Schmidt, B. & Loßnitzer, T. (2010). Lehrveranstaltungsevaluation: State of the Art, ein Definitionsvorschlag und Entwicklungslinien. *Zeitschrift für Evaluation*, *9*, 49 – 72.

Schreiber, J. B., Stage, F. K., King, J., Nora, A. & Barlow, E. A. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research*, *99*, 323 – 337.

Skinner, C. J., Holt, D. & Smith, T. M. F. (Eds.). (1989). *Analysis of complex surveys.* West Sussex, England: Wiley.

Spooren, P., Brockx, B. & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research*, *83*, 598 – 642.

Toland, M. D. & de Ayala, R. J. (2005). A multilevel factor analysis of students' evaluations of teaching. *Educational and Psychological Measurement*, *65*, 272 – 296.

Wu, J. & Kwok, O. (2012). Using SEM to analyze complex survey data: A comparison between design-based single-level and model-based multilevel approaches. *Structural Equation Modeling*, *19*, 16 – 35.

Dipl.-Psych. Erik Sengewald

Bundesagentur für Arbeit
GS 422 BPS
Regensburger Str. 104
90478 Nürnberg
E-Mail: erik@sengewald.net; Erik.Sengewald@Arbeitsagentur.de

Dipl.-Psych. Anja Vetterlein

Friedrich-Schiller-Universität Jena
Institut für Psychologie
Am Steiger 3, Haus 1
07743 Jena

5 Manuskript 2

Einfluss der Mehrfachevaluation auf die Ergebnisse konfirmatorischer Faktorenanalysen in der Lehrveranstaltungsevaluation

Erik Sengewald
Psychologische Forschung und Entwicklung
Bundesagentur für Arbeit

Anja Vetterlein
Institut für Psychologie
Friedrich-Schiller-Universität Jena

Zusammenfassung

Die Evaluation von Lehrveranstaltungen ist ein gängiges Feedbackinstrument an Universitäten. Während eines Semesters evaluieren Studenten mehr als eine Veranstaltung desselben oder unterschiedlicher Dozenten. Führt man die veranstaltungsspezifischen Stichproben zusammen, realisiert sich durch diese Mehrfachevaluation und die Gruppierung von Studenten zu Veranstaltungen eine hierarchische Struktur in der Gesamtstichprobe. Dennoch werden, zum Beispiel bei konfirmatorischen Faktorenanalysen (CFA) zur Überprüfung des Messmodells des eingesetzten Fragebogens, diese strukturellen Eigenschaften der Stichprobe oft ignoriert (Rindermann, 2009; Toland & de Ayala, 2005). Welchen Einfluss die Mehrfachevaluation auf die Modellgeltungskontrolle bei Fragebögen zur LVE hat ist bisher unbekannt. Durch die Anwendung von vier verschiedenen Stichprobentypen werden in der vorliegenden Studie vier unterschiedliche Varianten der Mehrfachevaluation in konkreten Stichproben realisiert und deren Einfluss auf die Modellgüte für verschiedene Verfahren der CFA untersucht. Die Ergebnisse zeigen, dass Mehrfachevaluationen auf Studenten- und Dozentenebene einen Einfluss auf die Modellpassung haben. Dieser Einfluss ist bei der CFA auf Veranstaltungsebene am größten und für die Multilevel-CFA am geringsten. Inwiefern sich die Mehrfachevaluation darüber hinaus auf die Stabilität geschätzter Faktorwerte niederschlägt, wird ebenfalls dargestellt.

Keywords: Mehrfachevaluation, Multiple-Membership-Modelle, Hochschulforschung, konfirmatorische Faktorenanalyse, Veranstaltungsqualität, Lehrevaluation

The effect of multiple evaluations on the results of confirmatory factor analysis of student evaluations of teaching

Abstract

Students' evaluation of teaching (SET) is a common method for feedback at universities. It is possible that students evaluate more than one course whithin a semester or during their studies. The aggregation of the course specific samples results in a hierarchical structure of the aggregated sample with multiple evaluations of students. However, this hierarchical structure of the data is seldom incorporated into confirmatory factor analysis (CFA) of the questionnaire used (Rindermann, 2009; Toland & de Ayala, 2005). Additionally nothing is known about the effect of multiple evaluations on the results of CFA for SETs. Accomplishing four different types of samples with varying amount of multiple evaluations, this study examines the effect of multiple evaluations on the results of different CFA methods. The results show that the specific structure of multiple evaluations differ in the effect on the results of a CFA. This effect, however, depends on the type of CFA used to analyse the data. A large effect appears, when CFA is done on course-level and only small effects appear when a multilevel CFA is used. Additionally, the effect of multiple evaluations on estimated factor scores is reported.

Keywords: multiple evaluations, multiple membership models, higher education, confirmatory factor analysis, course quality, students' evaluation of teaching

Kontakt: Erik Sengewald, E-Mail: erik@sengewald.net. Dieser Artikel wurde im Rahmen des gemeinsamen Bund-Länder-Programms für bessere Studienbedingungen und mehr Qualität in der Lehre aus Mitteln des Bundesministeriums für Bildung und Forschung (BMBF) unter dem Förderkennzeichen 01PL12071 gefördert.

Einleitung

Die Lehrveranstaltungsevaluation (LVE) wird als Feedbackinstrument, zur Steuerung im Rahmen der Qualitätsentwicklung und zur Untersuchung weiterführender wissenschaftlicher Fragestellungen eingesetzt. Die Verwendung der LVE als Feedbackinstrument für den Dozenten einer Veranstaltung ist weit verbreitet (vgl. Loßnitzer, Schmidt & Born, 2007; Marsh, 2007; Rindermann, 2009; Vetterlein & Sengewald, 2015). Hierfür werden die gewonnenen Evaluationsergebnisse deskriptiv ausgewertet und in einem Ergebnisbericht zusammengestellt. Auf Grundlage der Antwortverteilung und der deskriptiven Kennwerte auf Itemebene, erhält der Dozent ein Feedback zur Qualität seiner Veranstaltung. Die Lehrveranstaltungsqualität (LVQ) ist darüber hinaus Gegenstand weiterführender Fragestellungen. Für die Analyse von Zusammenhängen zwischen Variablen des Dozenten und der LVQ werden mehrere veranstaltungsspezifische Stichproben benötigt. Auch für konfirmatorische Faktorenanalysen (CFA) von Fragebögen zur LVE werden große Stichproben benötigt, die mehr als eine Veranstaltung beinhalten. Mit Hilfe der CFA wird untersucht, inwiefern das theoretische Messmodell des LVE-Instruments zur Messung der Qualität einer Lehrveranstaltung zu den konkreten Daten passt. Eine Passung des Messmodells ist die notwendige Voraussetzung für die Verwendung der LVE-Ergebnisse als Kriterium in weiterführenden Untersuchungen.

Diese veranstaltungsübergreifenden Untersuchungen basieren nur selten auf separat durchgeführten Erhebungen zur LVQ, vielmehr werden bestehende veranstaltungsspezifische Evaluationsergebnisse zu einer Gesamtstichprobe zusammengeführt. Dadurch können mehrere Veranstaltungen eines Dozenten (Mehrfachevaluation auf Dozentenebene) und mehrere Evaluationen derselben Studenten (Mehrfachevaluation auf Studentenebene) in der Gesamtstichprobe enthalten sein. Diese charakteristische Datenstruktur wird bei der Analyse der Messmodelle von Fragebögen zur LVE oft vernachlässigt (Toland & de Ayala, 2005). In diesem Zusammenhang wird auch das konkrete Verfahren der konfirmatorischen Faktorenanalyse (CFA-Verfahren) nur selten kritisch reflektiert. Die vorrangig eingesetzten CFA-Verfahren zur Überprüfung des Messmodells basieren auf Studentenebene unter Verwendung der Rohdaten oder auf Veranstaltungsebene nach Aggregation der Werte innerhalb einer Veranstaltung zu itemspezifischen Veranstaltungsmittelwerten.

Diese CFA-Verfahren zeigen für den Großteil der LVE-Instrumente keine zufriedenstellende Modellpassungen (Marsh et al., 2009). Das verwendete CFA-Verfahren oder die Mehrfachevaluation werden dabei selten als mögliche Ursache für den schlechten Modellfit in Betracht gezogen. Um dennoch den Modellfit zu verbessern, werden verschiedene Formen der exploratorischen Faktorenanalysen (EFA; vlg. Rindermann, 2009; Schmidt & Loßnitzer, 2010; Spooren, Brockx & Mortelmans, 2013) oder eine Kombination aus EFA und Strukturgleichungsmodellen (ESEM; Marsh et al., 2009) eingesetzt. Auch grundlegende Veränderungen des Itemmaterials werden häufig in Erwägung gezogen, um eine Verbesserung des Modellfits zu erreichen. Die konfirmatorische Mulitlevel-Analyse (ML-CFA) wird im Rahmen der LVE bisher selten eingesetzt (vgl. Sengewald & Vetterlein, 2015; Toland & de Ayala, 2005). Sengewald und Vetterlein (2015) vergleichen die drei Ansätze der CFA zur Überprüfung des Messmodells eines LVE-Fragebogens am Beispiel des Instruments PELVE (Prozess- und Ergebnisorientierte Lehrveranstaltungsevaluation; Loßnitzer et al., 2007) und verdeutlichen die Überlegenheit der ML-CFA gegenüber den konventionellen Verfahren. Die ML-CFA zeigt deutlich bessere Kennwerte zur Modellpassung und repräsentiert das theoretische Messmodell durch die Konstruktion latenter Variablen auf Studenten- und Veranstaltungsebene am besten.

Die Konsequenzen der Mehrfachevaluation für die Ergebnisse der Faktorenanalysen sind im Rahmen der LVE weder für konventionelle CFA-Verfahren noch für die ML-CFA untersucht. Toland und de Ayala (2005) beachten den Aspekt der Mehrfachevaluation explizit im Design der Studie, indem sie nur je eine Evaluation eines Dozenten und Studenten zulassen. Sengewald und Vetterlein (2015) untersuchen die Mehrfachevaluation nicht direkt, deuten jedoch unter Verwendung einer Minimalstichprobe, die aufgrund ihrer geringen Stichprobengröße nur ein geringfügiges Maß an Mehrfachevaluation aufweist, bereits auf das Problem hin.

Die vorliegende Studie verdeutlicht das Problem der Mehrfachevaluation zunächst anhand verschiedener Netzwerkgrafiken und ordnet es in den Kontext der Multilevel-Multiple-Membership-Modelle (MMMM) ein. Anschließend werden am Beispiel der LVE an der Friedrich-Schiller-Universität Jena (FSU Jena) die Auswirkungen der Missach-

tung von Mehrfachevaluationen auf die Ergebnisse der konfirmatorischen Faktorenanalyse des eingesetzten Fragebogens (PELVE) untersucht. Im Vordergrund steht dabei die Frage der Passung des theoretischen Messmodells zu konkreten empirischen Stichproben, die in ihrer Zusammensetzung variieren und Mehrfachevaluation auf Dozenten- und/oder Studentenebene enthalten.

Theorie

Zur systematischen Untersuchung des Einflusses der Mehrfachevaluation auf die CFA-Ergebnisse werden die typischen CFA-Verfahren im Methodenteil vorgestellt und die verschiedenen Datenstrukturen in LVE-Stichproben (Stichprobentypen) erläutert. In Abhängigkeit vom Studiendesign, das der Überprüfung des Messmodells zugrunde liegt und für die Stichprobenziehung ausschlaggebend ist, können Stichproben mit bzw. ohne Mehrfachevaluation auf Dozenten- und Studentenbene (mDmS bzw. oDoS), mit Mehrfachevaluation auf Dozentenebene (mDoS) und mit Mehrfachevaluation auf Studentenebene (oDmS) realisiert werden. Die Struktur typischer Stichproben in der LVE mit Fokus auf die Mehrfachevaluation wird im Folgenden genauer erläutert.

Mit Mehrfachevaluation auf Dozenten- und Studentenebene (mDmS). Ausgangspunkt für die CFA ist oft eine Gesamtstichprobe, die nach Zusammenführen veranstaltungsspezifischer Stichproben entsteht. Die Gesamtstichprobe kann Evaluationen verschiedener Veranstaltungen eines Dozenten (Mehrfachevaluation auf Dozentenebene) und mehrere Evaluationen eines Studenten enthalten (Mehrfachevaluation auf Studentenebene). Abbildung 1 verdeutlicht die Struktur, die in der resultierenden Stichprobe vorliegen kann. Es handelt sich um ein Multiple-Membership-Multilevel-Modell mit drei Ebenen (siehe auch Chung & Beretvas, 2012; Fielding & Goldstein, 2006), wobei Studenten in Veranstaltungen gruppiert sind. Eine Multilevel-Struktur mit zwei Ebenen wurde bereits in der konfirmatorischen Multilevel Analyse (ML-CFA) von Sengewald und Vetterlein (2015) berücksichtigt. Werden pro Dozent mehrere Veranstaltungen evaluiert, dann sind diese Veranstaltungen innerhalb der Dozenten gruppiert und eine Multilevel-Struktur mit drei Ebenen liegt vor. Die exemplarisch gewählten Dozenten A und B (vgl. Abbildung 1) haben jeweils drei Veranstaltungen evaluieren lassen. Darüber hinaus ist es plausibel anzunehmen, dass die Ver-

anstaltungen nicht nur unterschiedliche Studenten enthalten. Auf Studentenebene entsteht durch die Stichprobenzusammenführung eine Multiple-Membership-Struktur, weil Studenten verschiedene Veranstaltungen desselben Dozenten (Student 3,6 und 12 in Abbildung 1) oder anderer Dozenten (Student 7, 8, 9, 10 und 14 in Abbildung 1) evaluieren.

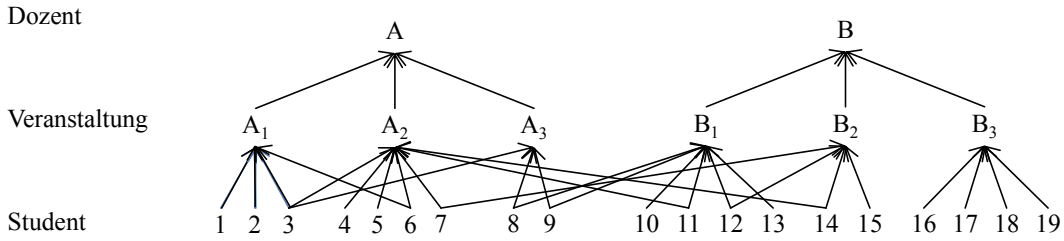


Abbildung 1. Netzwerkgrafik einer Stichprobe mit Mehrfachevaluation auf Dozenten- und Studentenebene.

Wie groß das jeweilige Ausmaß der Mehrfachevaluation im konkreten Fall ist, kann stark variieren. Außerdem ist das Ausmaß der Mehrfachevaluation in der LVE im Allgemeinen unbekannt, weil die Studenten nicht identifizierbar sind. Um die Mehrfachevaluation auf Studenten- und Dozentenebene zu erkennen, ist eine eindeutige Identifikation der Studenten und Dozenten notwendig. Erst dann kann die in Abbildung 1 dargestellte Struktur entdeckt und ggf. bei der CFA berücksichtigt werden.

Mehrfachevaluation auf Studentenebene (oDmS). Eine Stichprobe des Typs oDmS liegt dann vor, wenn je Dozent nur eine Veranstaltung in die Stichprobe zur CFA eingeht. Auf Studentenebene kann jedoch weiterhin Mehrfachevaluation enthalten sein. Dabei handelt es sich um Evaluationen von Veranstaltungen unterschiedlicher Dozenten durch dieselben Studenten. Dass Studenten verschiedene Veranstaltungen desselben Dozenten bewerten, ist bei diesem Stichprobentyp bzw. Erhebungsdesign ausgeschlossen. Im vorliegenden Beispiel von Dozent A und B (vgl. Abbildung 1) kann zum Beispiel die Veranstaltung A2 von Dozent A und B2 von Dozent B in die Gesamtstichprobe eingehen. Auf Studentenebene evaluieren Student 7 und 14 beide Veranstaltungen (vgl. Abbildung 2). Der Stichprobentyp lässt sich, sofern eine Dozenten-ID erhoben wird, nachträglich durch zufällige Auswahl einer Veranstaltung des Dozenten aus allen, die von ihm evaluiert wurden, realisieren. Der Vorteil dieser Methode liegt in ihrer Wiederholbarkeit, sodass beliebig viele konkrete Stichproben aus der Gesamtstichprobe gezogen werden können und zufällige Effekte einer einmaligen Realisie-

rung vermieden werden können.

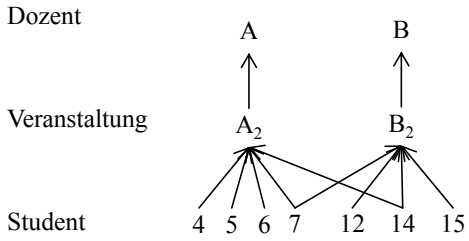


Abbildung 2. Netzwerkgrafik einer Stichprobe mit Mehrfachevaluation auf Studentenebene.

Durch die Evaluation mehrerer Veranstaltungen durch denselben Studenten können Abhängigkeiten in den Daten vorliegen, die durch den Studenten zustande kommen. Die i.i.d.-Annahme kann demnach nur innerhalb, nicht aber über mehrere Veranstaltungen hinweg, aufrechterhalten werden (vgl. Skinner, Holt & Smith, 1989). Die Evaluationsergebnisse der Veranstaltungen A2 und B2 (vgl. Abbildung 2) sollten zwar unabhängig voneinander sein, sind es jedoch möglicherweise nicht, weil zum Teil die gleichen Studierenden die Veranstaltungen evaluieren. Dies kann sich auf die Varianzen und Kovarianzen der Items und damit auf die Ergebnisse der CFA auswirken, wenn sich dadurch die Abweichungen der empirischen zur modellimplizierten Varianz-Kovarianz-Matrix ändern. Untersuchungen im Rahmen der Bildungsforschung zeigen, dass die Zugehörigkeit eines Schülers zu mehreren Schulen verzerrte Ergebnisse liefert, sofern diese Multiple-Membership-Struktur nicht berücksichtigt wird (Chung & Beretvas, 2012). Für die LVE liegt eine vergleichbare Situation vor.

Stichprobe ohne Mehrfachevaluationen (oDoS). Toland und de Ayala (2005) konnten in ihrer Studie eine gezielte Stichprobe erheben in der weder Dozenten noch Studenten mehrfach vorkamen. Hierfür wird nicht nur eine Veranstaltung je Dozent ausgewählt, sondern auch auf Studentenebene wird nur eine Evaluation zugelassen. Ein Untersuchungsdesign, das aus der prä-facto-Perspektive eine zufällige Ziehung von nur einer Evaluation der Studenten fordert, ist in der Praxis der LVE nur schwer umsetzbar. Einfacher zu realisieren ist hingegen, zunächst alle Teilnehmer einer zufällig gezogenen Veranstaltung an der Evaluation teilnehmen zu lassen und im Anschluss unter den mehrfach evaluierenden Studenten zufällig nur eine Evaluation in der Stichprobe zu belassen. Abbildung 3 zeigt hierfür ein Bei-

spiel. Es wurde auf Basis der Stichprobe aus Abbildung 2 je Student zufällig eine Evaluation ausgewählt. Für Student 7 wurde die Evaluation der Veranstaltung A2 und für Student 14 die Evaluation der Veranstaltung B2 ausgewählt (vgl. Abbildung 3).

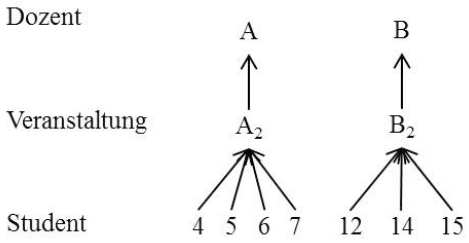


Abbildung 3. Netzwerkgrafik einer Stichprobe ohne Mehrfachevaluation auf Dozenten- und Studentenebene.

Mehrfachevaluation auf Dozentenebene (mDoS). Mehrfachevaluation auf Dozentenebene und ohne Mehrfachevaluation auf Studentenebene sind in der LVE selten. Das Design zur Erhebung eines derartigen Stichprobentyps müsste, ähnlich wie oben beschrieben, prä-facto die Evaluationen je Student kennen, um je Student nur eine zufällig ziehen zu können. Ähnlich wie für den Stichprobentyp oDoS, ist die Realisation einer mDoS Stichprobe post-hoc leichter durchführbar. Abbildung 4 zeigt ein Beispiel für die Realisierung einer mDoS-Stichprobe. Es sind alle Veranstaltungen der Dozenten A und B (vgl. Abbildung 1) und je Student genau eine Evaluation enthalten. Für alle Studenten ist nur eine Evaluation enthalten.

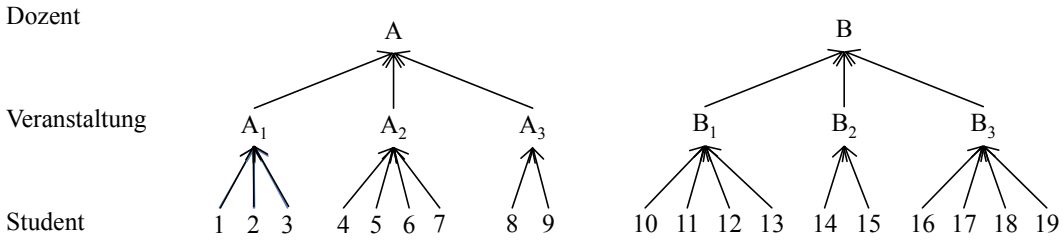


Abbildung 4. Netzwerkgrafik einer Stichprobe mit Mehrfachevaluation auf Dozentenebene.

Stichproben mit Mehrfachevaluation auf Dozentenebene können durch ein Multilevel-Modell mit drei Ebenen beschrieben werden. Im Gegensatz zu den Stichproben oDoS und oDmS, kann zwischen der Dozenten- und Veranstaltungsebene unterschieden werden. Beide Ebenen können eine Varianzquelle für die Evaluationsergebnisse sein. So kann sich zum Beispiel die Varianz der LVE-Ergebnisse zwischen Veranstaltungen gleicher

und verschiedener Dozenten unterscheiden. Inwiefern eine Vernachlässigung der dritten Ebene (Mehrfachevaluation auf Dozentenebene) für die Analyse von LVE-Ergebnissen von Bedeutung ist, wird in der vorliegenden Studie untersucht.

Fragebogen und Messmodell. Die vorliegende Studie verwendet LVE-Ergebnisse, die mit dem Fragebogen zur Prozess- und Ergebnisorientierten Lehrveranstaltungsevaluation (PELVE, vgl. Loßnitzer et al., 2007; Sengewald & Vetterlein, 2015) erhoben wurden. Der PELVE erfüllt die Kriterien der Definition von Lehrveranstaltungsqualität nach Schmidt und Loßnitzer (2010) und wird an der FSU Jena seit 2004 für die LVE eingesetzt. Insgesamt umfasst der Fragebogen 35 veranstaltungsübergreifende Ratingitems, die durch spezifische Items für Vorlesungen, Seminare und Übungen ergänzt werden. Demographische Angaben, Items zu Arbeitsaufwand, freitextlichen Anmerkungen und optionale Items vervollständigen den Fragebogen. Items der Kompetenzdimensionen werden auf einer fünfstufigen Likert-Skala mit den Antwortpolen *wenig* bis *viel* beantwortet. Alle anderen Ratingitems erfragen den Grad der Zustimmung auf einer fünfstufigen Antwortskala von *stimme nicht zu* bis *stimme zu*. Die Items sollen verschiedene Dimensionen der LVQ erfassen, die sich in Prozess- und Ergebnisvariablen unterteilen lassen. Sengewald und Vetterlein (2015) zeigen, dass für den PELVE ein multidimensionales Multilevel-Messmodell mit akzeptabler Passung vorliegt. Das Modell postuliert sieben Dimensionen der Veranstaltungsqualität und beinhaltet 33 Items mit Einfachladung auf je einer der sieben Dimensionen. Das Multilevel-Messmodell des PELVE ist schematisch in Abbildung 5 dargestellt.

Fragestellung. In der vorliegenden Studie werden die Auswirkungen der Mehrfachevaluation auf die Ergebnisse verschiedener CFA-Verfahren zur Überprüfung des Messmodells eines Fragebogens zur LVE untersucht. Dabei steht die Frage im Vordergrund, welche Unterschiede zwischen den CFA-Ergebnisen vorzufinden sind, wenn unterschiedliche Stichprobentypen mit bzw. ohne Mehrfachevaluationen verwendet werden. Hierfür werden auf Basis einer Gesamtstichprobe mit allen seit 2005 evaluierten Lehrveranstaltungen an der FSU Jena Stichproben gezogen, die eine Mehrfachevaluation auf Studenten- oder/und Dozentenebene beinhalten oder nicht beinhalten. Sofern die Mehrfachevaluation einen Einfluss auf die Kennwerte des Modellfits hat, und damit auf die Beurteilung der Modellgüte,

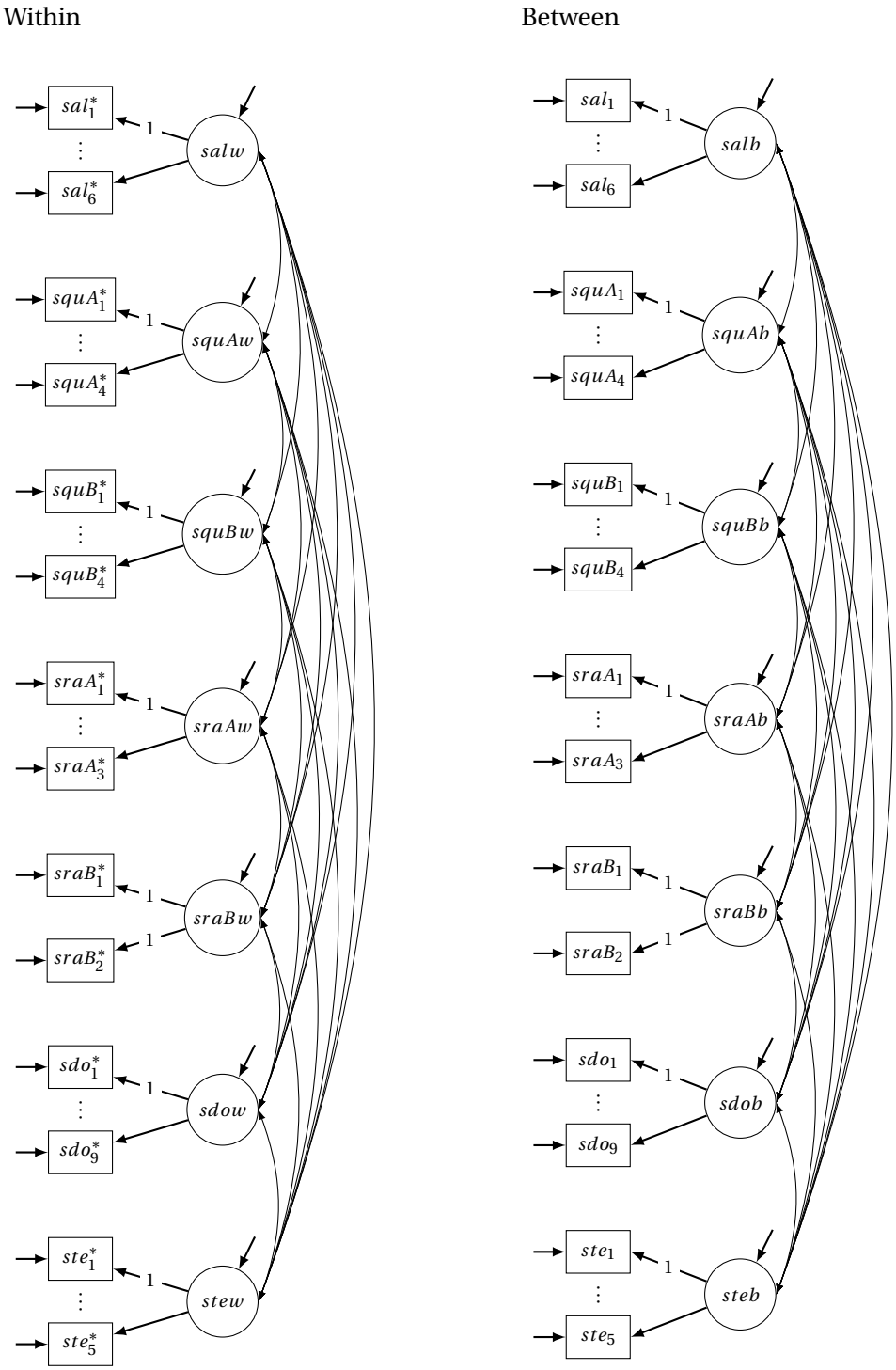


Abbildung 5. Schematische Darstellung des Multilevel-Messmodells des PELVE mit Studenten- (Within) und Veranstaltungsebene (Between).
Anmerkungen: Gesamteindruck (*sal*), Fachkompetenz (*squA*), sonstige Kompetenzen (*squB*), Rahmenbedingungen (*sraA*), Begleitmaterialien (*sraB*), Dozentenverhalten (*sdo*) und Studentenverhalten (*ste*). Diese Abkürzungen in Kombination mit *w* bzw. *b* verdeutlichen die Modellebene (within bzw. between). Die numerischen Subskripts repräsentieren die Itemnummer und das Symbol * die latent response Variable.

können Implikationen für die Überprüfung der Messmodelle von Fragebögen zur LVE abgeleitet werden. Dabei werden in der vorliegenden Studie die drei in der LVE vorherrschenden CFA-Verfahren (CFA auf Studentenebene, CFA auf Veranstaltungsebene und Multilevel-CFA) berücksichtigt und der Einfluss der Mehrfachevaluation für jedes CFA-Verfahren separat betrachtet.

Methode

Stichprobenziehung. In der vorliegenden Studie werden LVE Ergebnisse an der FSU Jena zu einer Gesamtstichprobe zusammengeführt, die als Datenbasis für die Überprüfung des Messmodells des PELVE dient. Durch die Verwendung eines Personencodes bzw. einer Dozenten-ID kann Mehrfachevaluation auf Studenten- bzw. Dozentenebene identifiziert und systematisch ausgeschlossen werden. Abbildung 6 zeigt das Vorgehen zur Realisation der verschiedenen Stichprobentypen (mDoS, oDmS und oDoS) aus der Gesamtstichprobe (mDmS).

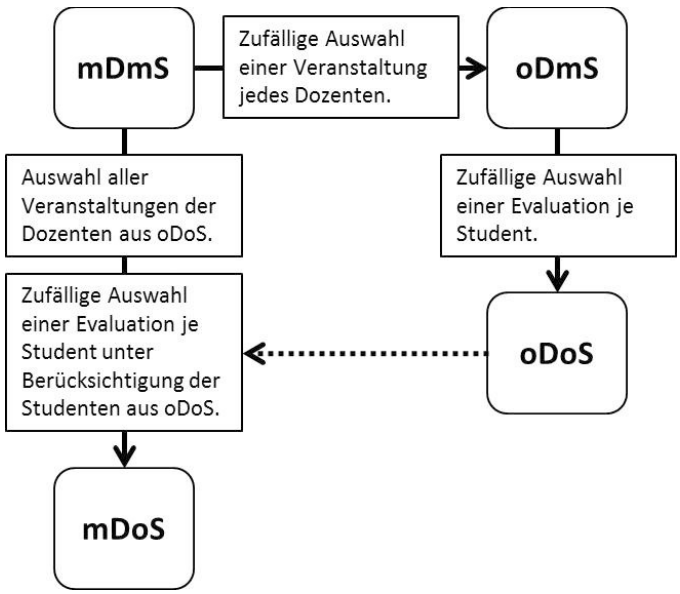


Abbildung 6. Verfahren zur Stichprobenziehung.
Anmerkungen: mDmS = mit Dozenten- mit Studentenwiederholung; mDoS = mit Dozenten- ohne Studentenwiederholung; oDmS = ohne Dozenten- mit Studentenwiederholung; oDoS = ohne Dozenten- ohne Studentenwiederholung.

Mit Hilfe einer Dozenten-ID kann die Mehrfachevaluation auf Dozentenebene ausgeschlossen werden. Auf Studentenebene kann die Mehrfachevaluation durch zufällige Auswahl von genau einer Evaluation je Student ausgeschlossen werden. Dafür wird ein

Personencode verwendet, um deren Angabe die Studenten in jedem Fragebogen gebeten werden. Der Personencode gewährleistet eine anonymisierte Identifikation von Evaluationen desselben Studenten. Dozenten-ID und Personencode sind ausreichend, um die in Abbildung 6 dargestellten Verfahren zur Generierung der Teilstichproben ohne Mehrfachevaluation auf Dozenten oder/und Studentenebene anzuwenden. Um Stichprobenfehler bei der Datenanalyse und der Interpretation zu minimieren, werden 100 Stichproben nach dem in Abbildung 6 beschriebenen Schema gezogen. Zunächst erfolgt die Ziehung der Stichprobe nach dem Modell oDmS durch zufällige Auswahl einer Veranstaltung jedes Dozenten. Anschließend wird diese Stichprobe um die Mehrfachevaluation auf Studentenebene bereinigt, indem zufällig eine Evaluation je Student gezogen wird. Die resultierende Stichprobe enthält weder Dozenten- noch Studentenwiederholung (oDoS). Die Stichprobe nach dem Modell oDoS ist zusammen mit der Gesamtstichprobe mDmS Grundlage für die Realisation der Stichproben nach dem Modell mDoS (vgl. Abbildung 6).

Mehrfachevaluation und Stichproben. In der vorliegenden Studie wurden alle LVE-Ergebnisse vom Sommersemester 2005 bis 2013 berücksichtigt, die von mindestens fünf Studenten evaluiert wurden. Die Stichprobenziehung und Datenaufbereitung erfolgten mit der Software R (R Core Team, 2014). Die resultierende Gesamtstichprobe mit $N_V = 7\,459$ Veranstaltungen von $N_D = 1\,603$ unterschiedlichen Dozenten mit insgesamt $N_S = 183\,334$ Studenten (vgl. Version *mDmS* in Tabelle 1) ist damit ausreichend groß, um eine CFA durchzuführen. In jeder LVE befinden sich mindestens fünf und durchschnittlich $N_{S\nu} = 24.58$ ($SD_{N_{S\nu}} = 25.89$) Studenten.

Die Gesamtstichprobe beinhaltet Mehrfachevaluation auf Studenten- und Dozentenebene. Pro Dozent liegen im Durchschnitt $\overline{N}_{Vd} = 3.45$ ($SD_{N_{Vd}} = 4.80$) und je Student $\overline{N}_{Vs} = 2.29$ ($SD_{N_{Vs}} = 2.26$) Evaluationen vor. Tabelle 1 zeigt die Verteilung der relevanten Kennwerte über die 100 konkreten Stichproben je Stichprobentyp. Alle Stichprobengrößen sind trotz der Reduktion durch die geforderten Restriktionen des jeweiligen Stichprobentyps groß genug, um CFA anzuwenden und verlässliche Aussagen bzgl. des Modellfits zu erhalten (vgl. Spalte N_V und N_S in Tabelle 1). Analog zur Studie von Chung

Tabelle 1
Gesamtstichprobe und realisierte Teilstichproben

Version	N_V	N_D	N_{Vd}	N_S	N_{Sv}	N_{Vs}
mDmS	7459 (–)	1603 (–)	3.5 (4.8)	183334 (–)	24.6 (26.0)	2.3 (2.3)
oDoS	788.5 (13.4)	788.5 (13.4)	1.00 (0.0)	8696.1 (0.0)	11.0 (0.2)	1.0 (0.0)
mDoS	2882.8 (65.5)	793.5 (14.7)	2.8 (0.1)	36317.8 (0.1)	12.6 (0.2)	1.0 (0.0)
oDmS	788.5 (13.4)	788.5 (13.4)	1.0 (0.0)	25438.8 (0.0)	32.3 (0.7)	1.2 (0.01)

Anmerkungen: Für die Stichproben, ohne Dozenten- und Studentenwiederholung (oDoS), mit Dozenten- und ohne Studentenwiederholung (mDoS) sowie für ohne Dozenten- und mit Studentenwiederholung (oDmS) sind Kennwerte dargestellt: Anzahl Veranstaltungen (N_V), Anzahl Dozenten (N_D), durchschnittliche Anzahl Veranstaltungen je Dozent (N_{Vd}), Anzahl Studenten (N_S), durchschnittliche Anzahl Studenten je Veranstaltung (N_{Sv}) und durchschnittliche Anzahl Veranstaltungen, die durch einen Studenten evaluiert werden (N_{Vs}). Die Tabelle enthält jeweils den Mittelwert (Standardabweichung) der Kennwerte für 100 Stichproben. Für den Stichprobentyp mit Dozenten- und Studentenwiederholung (mDmS) liegt nur eine Stichprobe vor.

und Beretvas (2012) zeigt Abbildung 7 die prozentuale Verteilung der Mehrfachevaluation. Hierdurch lässt sich der Anteil mehrfach evaluierender Studenten und Dozenten besser beurteilen als durch die Betrachtung des Mittelwerts. Im Gegensatz zu Untersuchungen zur Multiple-Membership-Struktur im Schulkontext (vgl. Chung & Beretvas, 2012), ist der Anteil an Studenten, die in mehr als zwei oder drei Gruppen sind, bei der LVE höher. Die empirische Verteilung in Abbildung 7 zeigt, dass eine hohe Variablilität bzgl. der Multiple-Membership-Struktur bei der LVE besteht. In der Gesamtstichprobe evaluieren 48.67 % der 183 334 Studenten mehr als eine Veranstaltung und 7.49 % mehr als zehn Veranstaltungen. Auf Dozentenebene lassen 61.76 % mehr als eine Veranstaltung evaluieren (vgl. mDmS in Abbildung 7). Abbildung 7 verdeutlicht zudem die Verteilung in den Teilstichproben. So enthält die Stichprobe mit Mehrfachevaluation auf Dozentenebene (mDoS) genau eine Veranstaltung pro Student (vgl. mDoS NVs in Abbildung 7) und im Durchschnitt 57.34 % Dozenten mit mehr als einer evaluierten Veranstaltung (vgl. mDoS NVd in Abbildung 7). In der Stichprobe oDmS ist die Mehrfachevaluation nur auf Studentenebene vorhanden, wobei 26.60 % mehr als eine Veranstaltung evaluieren. Eine Eingruppierung von über einem Viertel der Studenten in mehreren Veranstaltungen kann sich dennoch auf den Modellfit auswirken, auch wenn die überwiegende Mehrheit nur einer Gruppe zugeordnet ist (vgl. Chung & Beretvas, 2012).

Die Varianz der Mehrfachevaluation ist in den Stichproben nach dem Modell *oDmS*

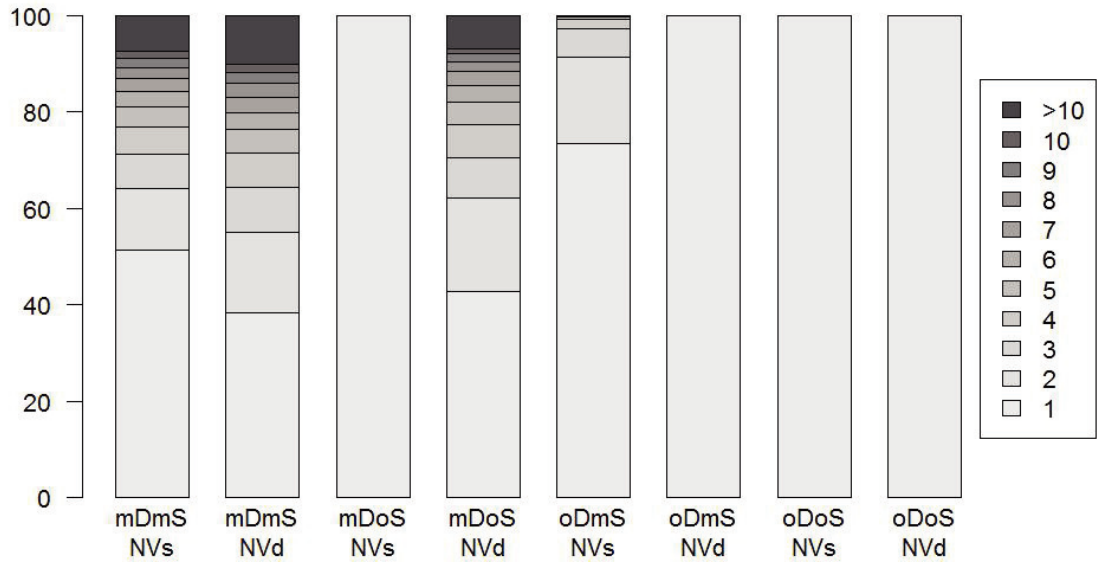


Abbildung 7. Verteilung des Anteils an Studenten und Dozenten mit Mehrfachevaluation. *Anmerkungen:* mDmS = mit Dozenten- und Studentenwiederholung, mDoS = mit Dozenten-, ohne Studentenwiederholung, oDmS = ohne Dozenten-, mit Studentenwiederholung, oDoS = ohne Dozenten- und Studentenwiederholung; NVs = Anzahl Veranstaltungen je Student, NVd = Anzahl Veranstaltungen je Dozent

deutlich geringer als in der Gesamtstichprobe und enthält nur sehr wenige Studenten, die mehr als drei Veranstaltungen evaluieren. Diese Situation ist gut mit der Simulationsstudie von Chung und Beretvas (2012) vergleichbar. Analog zur Tabelle 1 zeigt Abbildung 7, dass im Stichprobentyp oDoS keinerlei Mehrfachevaluationen enthalten sind und 100 % der Dozenten und Studenten mit nur einer Evaluation in die Stichproben eingehen.

Konfirmatorische Faktorenanalyse (CFA). Zur Prüfung des Messmodells auf Studentenebene wird eine CFA für ordinale Variablen angewandt. Nach B. O. Muthén und Asparouhov (in Druck) werden zunächst I kontinuierliche Latent-Response-Variablen (LRV) Y^*_{pi} $(i = 1,2,...,I)$ angenommen, die einem linearen Messmodell mit D latenten Variablen ϑ_{pd} folgen (vgl. Gleichung 1). Der Zusammenhang zwischen manifester Variable Y_{pi} und LRV Y^*_{pi} wird durch ein Schwellenmodell beschrieben (vgl. B. O. Muthén, 1984; B. O. Muthén & Asparouhov, in Druck). Für $A = 5$ kategoriale Items ist das allgemeine Schwellenmodell mit $A-1$ Schwellen in Gleichung 2 dargestellt. Die manifeste Variable Y_{pi} nimmt einen ihrer A Werte an, wenn Y^*_{pi} eine bestimmte Schwelle τ_{ia} $(a = 1,2,...,A-1)$ über- bzw. unterschreitet (vgl. Gleichung 2).

$$Y^*_{pi} = \nu_i + \sum_{d=1}^{D} \lambda_{id}\vartheta_{pd} + \varepsilon_{pi} \tag{1}$$

$$Y_{pi} = \begin{cases} 0, & \text{wenn} \quad Y^*_{pi} \leq \tau_{i1} \\ 1, & \text{wenn} \quad \tau_{i1} < Y^*_{pi} \leq \tau_{i2} \\ 2, & \text{wenn} \quad \tau_{i2} < Y^*_{pi} \leq \tau_{i3} \\ 3, & \text{wenn} \quad \tau_{i3} < Y^*_{pi} \leq \tau_{i4} \\ 4, & \text{wenn} \quad \tau_{i4} < Y^*_{pi} \end{cases} \tag{2}$$

Gleichung 1 verdeutlicht, dass die Gruppierung der Studenten in Veranstaltungen vernachlässigt wird. Die latenten Variablen ϑ_{pd} sind auf Personenebene und nicht auf Veranstaltungsebene definiert (vgl. Gleichung 1).

Um dies zu ermöglichen, wird die CFA auf Veranstaltungsebene empfohlen (vgl. z. B. Clayson, 2007; Marsh, 1983; Marsh & Roche, 1997; Rindermann, 2009). Die CFA auf Veranstaltungsebene beruht auf den Mittelwerten der I Items in J Veranstaltungen. Bezüglich der Variablen $\overline{Y}_{ij}$ wird ein lineares Messmodell mit D latenten Variablen ϑ_{dj} aufgestellt (vgl. Gleichung 3). Der Veranstaltungsmittelwert $\overline{Y}_{ij}$ eines Items i in der Veranstaltung j wird über den Mittelwert aller N_j Personen in der Veranstaltung j errechnet (vgl. Gleichung 4). Die individuellen Urteile eines Studenten p in Veranstaltung j auf einem Item i (y_{pij}) werden im Messmodell selbst nicht berücksichtigt (vgl. Gleichung 3).

$$\overline{Y}_{ij} = \nu_i + \sum_{d=1}^{D} \lambda_{id} \cdot \vartheta_{dj} + \varepsilon_{ij} \tag{3}$$

$$\text{mit} \quad \overline{Y}_{ij} = \frac{1}{N_j} \cdot \sum_{p=1}^{N_j} Y_{pij} \tag{4}$$

Durch die Aggregation zu veranstaltungsspezifischen Itemmittelwerten nimmt man an, dass Effekte der Mehrfachevaluation ausgeschlossen werden können (Rindermann, 2009).

Ein weiteres CFA-Verfahren ist die ML-CFA, mit der es möglich ist, unter Verwendung der Erhebungseinheit (Studentenebene), Aussagen über latente Variablen auf der interessierenden Analyseebene (Veranstaltungsebene) zu treffen. Dafür wird ein Multilevel-Messmodell mit Studentenebene (Within-Messmodell) und Veranstaltungsebene (Between-Messmodell) erstellt (vgl. Asparouhov & Muthén, 2006; B. O. Muthén, 1994; B. O. Muthén & Asparouhov, in Druck; Toland & de Ayala, 2005). Basierend auf dem Schwellenmodell (Gleichung 2) wird ein Messmodell für die latenten Variablen auf Within-Ebene (ϑ_{Wd}) aufgestellt (vgl. Gleichung 5). Der Index W signalisiert, dass es sich um ebenenspezifische latente Variablen (ϑ_{Wd}), Ladungen (λ_{Wid}) und Residuen (ε_{Wij}) innerhalb einer Veranstaltung (Within) handelt. Für das Intercept ν_{ij} wird ein lineares Between-Messmodell aufgestellt (vgl. Gleichung 6). In diesem werden latente Variablen auf Veranstaltungsebene (ϑ_{Bdj}) definiert. Der Index B kennzeichnet hier die Between-Ebene (Veranstaltungsebene). Das Intercept ν_{ij} aus Gleichung 5 variiert demnach in Abhängigkeit des Wertes der latenten Variablen ϑ_{Bdj} der Veranstaltung j auf der Dimension d. Das explizit formulierte Within- und Between-Messmodell wird im Rahmen der Modellgeltungskontrolle der ML-CFA geprüft.

$$\text{Within} \qquad Y^*_{ij} = \nu_{ij} + \sum_{d=1}^{D_W} \lambda_{Wid}\vartheta_{Wd} + \varepsilon_{Wij} \tag{5}$$

$$\text{Between} \qquad \nu_{ij} = \nu_i + \sum_{d=1}^{D_B} \lambda_{Bid}\vartheta_{Bjd} + \varepsilon_{Bij} \tag{6}$$

Die oben beschriebenen CFA-Verfahren (CFA auf Studenteneben, CFA auf Veranstaltungsebene und ML-CFA) werden unabhängig vom Stichprobentyp für jede Stichprobe durchgeführt. Für jeden Kennwert des Modellfits wird eine Verteilung auf Basis der CFA-Ergebnisse aus den 100 Stichproben erstellt. Durch die Analyse an identischen Stichproben sind die Verteilungen zwischen den CFA-Verfahren vergleichbar. Innerhalb jedes CFA-Verfahrens sind die Ergebnisse für die unterschiedlichen Stichprobentypen vergleichbar. Damit lässt sich der Einfluss der Mehrfachevaluation auf die CFA-Ergebnisse innerhalb eines CFA-Verfahrens vergleichen.

Alle Varianten der CFA wurden mit der Software Mplus (L. K. Muthén & Muthén, 2008) berechnet. Durch die hohe Stichprobenabhängigkeit des χ^2-Wertes im Rahmen des Mo-

delltests werden deskriptive Fitindizes wie der RMSEA, CFI, SRMR und das Verhältnis aus χ^2-Wert und Freiheitsgraden (*df*) zur Beurteilung der Modellgüte empfohlen (vgl. Browne & Cudeck, 1993; Schermelleh-Engel, Moosbrugger & Müller, 2003). Der RMSEA wird bei Werten RMSEA ≤ .05 als gut, für Werte .05 < RMSEA ≤ .08 als akzeptabel, für Werte .08 < RMSEA ≤ .10 als mittelmäßig und für die Werte .10 < RMSEA als inakzeptabel bewertet (vgl. Browne & Cudeck, 1993). Für den CFI können Werte über .97 als gut bewertet werden, Werte über .95 sind noch akzeptabel (Schermelleh-Engel et al., 2003). Der SRMR ist ebenfalls sensitiv gegenüber Missspezifikation des Messmodells und nur geringfügig von der Stichprobengröße abhängig (Hu & Bentler, 1998). Bei einem guten Modell sollte der SRMR kleiner als .05 und für ein akzeptables Modell kleiner als .10 sein (Hu & Bentler, 1995). Eine getrennte Beurteilung des Within- und Between-Messmodells ist durch den SRMR möglich. Die Simulationsstudie von Ryu und West (2009) zeigt, dass sowohl RMSEA als auch der CFI ungeeignet sind und der SRMR-Between verwendet werden soll. Die Angabe des χ^2/df-Verhältnisses kann für alle CFA-Verfahren angegeben werden. Dieser Kennwert kann jedoch nur zum Vergleich verschiedener Modelle bei gleicher Stichprobe herangezogen werden. Vergleicht man verschiedene Stichproben, wie es bei der Fragestellung nach dem Einfluss der Mehrfachevaluation der Fall ist, ist dieser Kennwert aufgrund seiner Sensitivität gegenüber der Stichprobengröße nur in Relation zur Stichprobengröße interpretierbar.

Faktorwerte. Unabhängig von der Beurteilung des Modellfits in Abhängigkeit der Mehrfachevaluation und des verwendeten CFA-Verfahrens, ist die Frage nach diagnostischen Implikationen bedeutsam. In der vorliegenden Studie werden hierfür die Faktorwerte der latenten Variablen geschätzt, die als Schätzer der LVQ verwendet werden können. Die Schätzung von Faktorwerten für die ML-CFA kann in Mplus durch den implementierten Bayes-Schätzer erfolgen. Asparouhov und Muthén (2010) konnten zeigen, dass der Bayes-Schätzer bessere Ergebnisse für kleine Stichproben liefert als ein ML-Verfahren. Bei großen Stichproben liefern beide Verfahren die gleichen Ergebnisse. Im Falle der CFA auf Studentenebene repräsentiert dieser Faktorwert den Wert der latenten Variablen auf Studentenebene. Für eine Interpretation auf Veranstaltungsebene wurde bei der CFA auf Studentenebene nachträglich der Mittelwert der Faktorwerte für jede Veranstaltung berechnet.

Im Falle der CFA auf Veranstaltungsebene und der ML-CFA liegen die geschätzten Faktorwerte direkt auf Ebene der Veranstaltung vor.

Mit Hilfe korrelativer Betrachtungen wird der Einfluss der Mehrfachevaluation auf die Rangreihe der jeweils geschätzten Dimension der Veranstaltungsqualität untersucht. Zusätzlich werden Prozentrangdifferenzen angegeben, die das Ausmaß der individuellen Veränderung in der Rangreihe darstellen. Bei der Interpretation von Prozentrangdifferenzen ist jedoch darauf zu achten, dass die Änderung des Rangplatzes einer Veranstaltung mit der Änderung des Rangplatzes einer anderen Veranstaltung einhergeht. Die Verteilung der Prozentrangdifferenzen kann daher nur deskriptiv in Relation zueinander interpretiert werden, sodass vor allem der Einfluss der Mehrfachevaluation auf die Rangreihe innerhalb eines CFA-Verfahrens untersucht werden kann.

Ergebnisse

Einfluss der Mehrfachevaluation auf den Modellfit. Für die je 100 Stichproben nach dem Modell mDoS, oDmS und oDoS liegen empirische Verteilungen der Kennwerte des Modellfits vor, die eine grafische Beurteilung der Unterschiede zulassen. Die Gesamtstichprobe (mDmS) wird durch je einen Wert für jedes Modellfit-Maß und CFA-Verfahren repräsentiert. Die χ^2-Modelltests sind alle signifikant ($p < .05$). Aufgrund der Abhängigkeit des χ^2-Wertes von der Stichprobengröße werden nur die deskriptiven Kennwerte des Modellfits zur Beurteilung der Modellgüte herangezogen. Der RMSEA in Abhängigkeit des CFA-Verfahrens und der Mehrfachevaluation ist in Abbildung 8 dargestellt. Für den CFI wird der Vergleich grafisch durch Abbildung 9 ermöglicht. Dargestellt sind jeweils Dichten der Kennwerte für die 100 Stichproben der Stichprobentypen mDoS, oDmS und oDoS. Die Gesamtstichprobe ist jeweils durch einen senkrechten Strich repräsentiert. Der Vergleich von Verteilungen innerhalb der CFA auf Studentenebene (CFA-S), der CFA auf Veranstaltungsebene (CFA-V) und der Multilevel-CFA (ML-CFA) ermöglicht die Beurteilung des Einflusses der Mehrfachevaluation auf den jeweiligen Kennwert (vgl. Abbildung 8 und Abbildung 9).

Wie bereits bei Sengewald und Vetterlein (2015), zeigt auch hier die ML-CFA die besten Kennwerte und die CFA auf Veranstaltungsebene liefert im Vergleich den schlechtesten

Modellfit. Eine Betrachtung des Einflusses der Mehrfachevaluation auf den Modellfit kann für jedes CFA-Verfahren separat erfolgen.

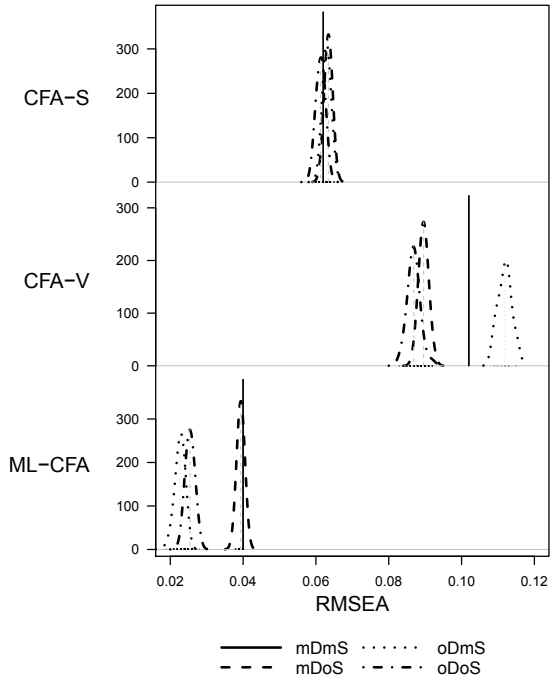


Abbildung 8. Dichte des RMSEA in Abhängigkeit der Mehrfachevaluation für die CFA auf Studentenebene (CFA-S), auf Veranstaltungsebene (CFA-V) und für die Multilevel-CFA (ML-CFA). Mehrfachevaluation: mDmS (mit Dozenten- und Studentenwiederholung), mDoS (mit Dozenten- ohne Studentenwiederholung), oDmS (ohne Dozenten- mit Studentenwiederholung), oDoS (ohne Dozenten- und Studentenwiederholung).

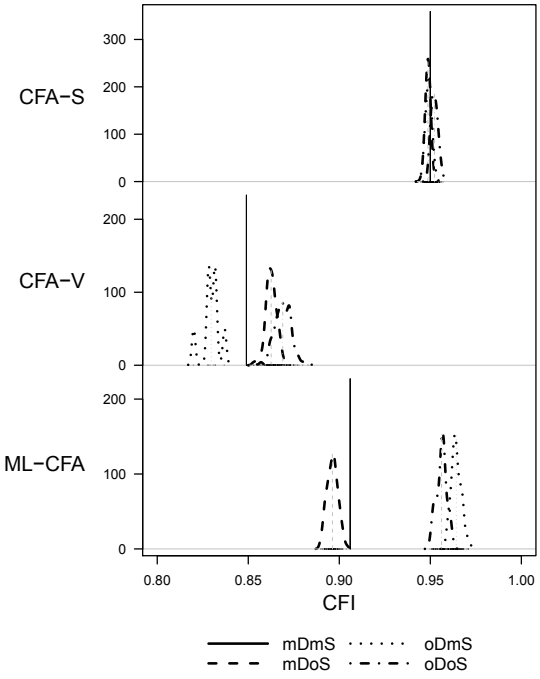


Abbildung 9. Dichte des CFI in Abhängigkeit der Mehrfachevaluation für die CFA auf Studentenebene (CFA-S), auf Veranstaltungsebene (CFA-V) und für die Multilevel-CFA (ML-CFA). Mehrfachevaluation: mDmS (mit Dozenten- und Studentenwiederholung), mDoS (mit Dozenten- ohne Studentenwiederholung), oDmS (ohne Dozenten- mit Studentenwiederholung), oDoS (ohne Dozenten- und Studentenwiederholung).

CFA auf Studentenebene. Die Kennwerte zur Beurteilung der Modellgüte für die CFA auf Studentenebene sind in Tabelle 2 in Abhängigkeit der Mehrfachevaluation dargestellt. RMSEA und CFI zeigen eine akzeptable Passung des Modells an. Für alle Stichprobentypen befinden sich die RMSEA-Werte zwischen .05 und .08, was auf einen akzeptablen Fit des Modells hindeutet (Browne & Cudeck, 1993). Die CFI-Werte liegen um den .95 Grenzwert und sind kleiner als .97. Damit können sie als akzeptabel interpretiert werden (vgl. Tabelle 2). Die Varianz dieser Kennwerte ist über die Stichproben hinweg nahe Null, sodass die stichprobenspezifischen Schwankungen vernachlässigbar sind. Die Unterschie-

de zwischen verschiedenen Bedingungen der Mehrfachevaluation sind ebenfalls nicht substantiell (vgl. Modell CFA-S in Abbildung 8 und Abbildung 9). Setzt man die χ^2/df-Werte aus Tabelle 2 ins Verhältnis zur durchschnittlichen Stichprobengröße (vgl. Tabelle 1), kann dieser Kennwert vergleichend eingesetzt werden. So liegt das Verhältnis aus χ^2/df-Wert und Stichprobengröße für den Stichprobentyp *oDoS* bei rund 0.0039 und für den Stichprobentyp mDoS bei rund 0.0040. Diese Werte zeigen keinen deutlichen Vorteil des Stichprobentyps *oDoS* gegenüber den anderen Stichprobentypen, wie es zunächst den Anschein bei der Beurteilung des χ^2/df-Wertes hat. Eine bessere Passung des Modells zu den empirischen Daten des Stichprobentyps oDoS im Vergleich zu den anderen Stichprobentypen kann für die CFA auf Studentenebene nicht attestiert werden. Die Mehrfachevaluation scheint hier keinen nennenswerten Einfluss auf die Kennwerte des Modellfits zu haben.

Tabelle 2
Verteilung der Fitmaße für CFA auf Studentenebene in Abhängigkeit der Mehrfachevaluation

Version	RMSEA	CFI	χ^2/df
mDmS	0.062 (–)	0.950 (–)	699.28 (–)
oDoS	0.061 (0.00)	0.952 (0.00)	33.63 (1.36)
mDoS	0.063 (0.00)	0.949 (0.00)	146.94 (3.39)
oDmS	0.063 (0.00)	0.949 (0.00)	102.09 (4.13)

Anmerkungen: Die Tabelle enthält die deskriptiven Fit-Maße RMSEA (mit RMSEA = $\frac{1}{N}\sum_{i=1}^{N}\text{RMSEA}_i$), CFI (mit CFI = $\frac{1}{N}\sum_{i=1}^{N}\text{CFI}_i$) und χ^2/df (mit $\chi^2/df = \frac{1}{N}\sum_{i=1}^{N}\chi_i^2/df_i$) für N=100 Stichproben je Version mit Dozenten- und Studentenwiederholung (mDmS), ohne Dozenten- und Studentenwiederholung (oDoS), mit Dozentenwiederholung (mDoS) und mit Studentenwiederholung (oDmS). In Klammern sind, falls verfügbar, die entsprechenden Standardabweichungen aufgeführt.

CFA auf Veranstaltungsebene. Der RMSEA und der CFI zeigen für die CFA auf Veranstaltungsebene deutlich schlechtere Werte als für die CFA auf Studentenebene (vgl. Modell CFA-V in Abbildung 8 und Abbildung 9). Die genauen Werte für den Modellfit können Tabelle 3 entnommen werden. Der RMSEA ist mit .087 auch für den Stichprobentyp *oDoS* und damit für den Fall ohne Mehrfachevaluationen nicht akzeptabel. Der Modellfit verschlechtert sich marginal, wenn Stichproben mit Mehrfachevaluation auf Dozentenebene

vorliegen. Eine deutliche Verschlechterung (um .025 auf RMSEA= .112) zeigt sich bei der Betrachtung der Stichproben mit Mehrfachevaluation auf Studentenebene (vgl. Tabelle 3).

Tabelle 3
Verteilung der Fitmaße für CFA auf Veranstaltungsebene in Abhängigkeit der Mehrfachevaluation

Version	RMSEA	CFI	χ^2/df
mDmS	0.102 (–)	0.849 (–)	77.99 (–)
oDoS	0.087 (0.00)	0.869 (0.00)	6.95 (0.25)
mDoS	0.090 (0.00)	0.863 (0.00)	24.12 (0.66)
oDmS	0.112 (0.00)	0.830 (0.00)	10.91 (0.45)

Anmerkungen: Die Tabelle enthält die deskriptiven Fit-Maße RMSEA (mit RMSEA = $\frac{1}{N}\sum_{i=1}^{N}\text{RMSEA}_i$), CFI (mit CFI = $\frac{1}{N}\sum_{i=1}^{N}\text{CFI}_i$) und χ^2/df (mit $\chi^2/df = \frac{1}{N}\sum_{i=1}^{N}\chi_i^2/df_i$) für N=100 Stichproben je Version mit Dozenten- und Studentenwiederholung (mDmS), ohne Dozenten- und Studentenwiederholung (oDoS), mit Dozentenwiederholung (mDoS) und mit Studentenwiederholung (oDmS). In Klammern sind, falls verfügbar, die entsprechenden Standardabweichungen aufgeführt.

Die Analyse des CFI zeigt ein ähnliches Bild (vgl. Tabelle 3 und Modell CFA-V in Abbildung 9). Auch hier ist der Modellfit als schlecht zu bewerten und der negative Einfluss der Mehrfachevaluation auf Studentenebene sichtbar. Für die bereinigte Stichprobe ohne Mehrfachevaluation (vgl. *oDoS* in Tabelle 3) verbessern sich die geschätzten Gütemaßstäbe, sie erreichen jedoch nicht die Grenzwerte eines guten Modells. Das gilt auch für die Betrachtung der χ^2/df-Werte im Verhältnis zur Stichprobengröße. Die Werte liegen zwischen 0.0084 für den Stichprobentyp *mDoS* und 0.014 für *oDmS*. Diese Werte sind für Stichprobentypen mit Mehrfachevaluation auf Studentenebene höher und damit schlechter als für Stichprobentypen ohne Mehrfachevaluation auf Studentenebene. Die Werte sind ebenfalls schlechter als diejenigen bei der CFA auf Studentenebene und unterstützen damit die Einschätzung auf Basis der RMSEA- und CFI-Verteilung. In der Literatur wird angenommen, dass mögliche Effekte der Mehrfachevaluation auf Studentenebene durch die vorgelagerte Aggregation zu Veranstaltungsmittelwerten beseitigt werden können (vgl. Rindermann, 2009). Diese Annahme kann durch die vorliegenden Ergebnisse jedoch widerlegt werden.

Multilevel-CFA. Die Analyse des Multilevel-Messmodells zeigt zunächst einen deutlich besseren Fit als die bisher vorgestellten Analysen (vgl. Tabelle 4 und Modell ML-CFA

Tabelle 4
Verteilung der Fitmaße für ML-CFA in Abhängigkeit der Mehrfachevaluation

Version	RMSEA	CFI	χ^2/df	SRMR_{within}	$\text{SRMR}_{between}$
mDmS	0.040 (–)	0.906 (–)	295.67 (–)	0.04 (–)	0.07 (–)
oDoS	0.025 (0.00)	0.956 (0.00)	6.64 (0.44)	0.04 (0.00)	0.08 (0.00)
mDoS	0.039 (0.00)	0.896 (0.00)	57.36 (1.89)	0.04 (0.00)	0.08 (0.00)
oDmS	0.023 (0.00)	0.964 (0.00)	14.09 (1.26)	0.04 (0.00)	0.08 (0.00)

Anmerkungen: Die Tabelle enthält die deskriptiven Fit-Maße RMSEA (mit $\text{RMSEA} = \frac{1}{N}\sum_{i=1}^{N}\text{RMSEA}_i$), CFI (mit $\text{CFI} = \frac{1}{N}\sum_{i=1}^{N}\text{CFI}_i$), χ^2/df (mit $\chi^2/df = \frac{1}{N}\sum_{i=1}^{N}\chi^2_i/df_i$), SRMR_{within} (mit $\text{SRMR}_{within} = \frac{1}{N}\sum_{i=1}^{N}\text{SRMR}_i^{within}$) und $\text{SRMR}_{between}$ (mit $\text{SRMR}_{between} = \frac{1}{N}\sum_{i=1}^{N}\text{SRMR}_i^{between}$) für N=100 Stichproben je Version mit Dozenten- und Studentenwiederholung (mDmS), ohne Dozenten- und Studentenwiederholung (oDoS), mit Dozentenwiederholung (mDoS) und mit Studentenwiederholung (oDmS). In Klammern sind, falls verfügbar, die entsprechenden Standardabweichungen aufgeführt.

in Abbildung 8 und Abbildung 9). Der RMSEA liegt für alle Stichprobentypen unter .05 und ist damit als gut zu bewerten. Die Veränderungen des RMSEA in Abhängigkeit der Mehrfachevaluation unterscheiden sich von den Ergebnissen der CFA auf Veranstaltungsebene. Im Vergleich zu Stichproben ohne Mehrfachevaluation (*oDoS*) verschlechtert sich der Modellfit bei der Analyse von Stichproben mit Mehrfachevaluation auf Dozentenebene (*mDoS* und *mDmS*). Bei der ML-CFA tritt jedoch keine Verschlechterung des RMSEA durch Mehrfachevaluation auf Studentenebene auf, wie es bei der CFA auf Veranstaltungsebene der Fall ist. Dieses Ergebnis findet sich auch bei der Analyse des Verhältnisses aus χ^2/df-Wert und Stichprobengröße wieder. Die Werte liegen zwischen 0.0006 für den Stichprobentyp oDmS und 0.0017 für den Stichprobentyp mDmS. Damit liegen alle Werte unter denen der CFA auf Studenten- und Veranstaltungsebene und sind für Stichprobentypen oDoS und oDmS deutlich geringer als für mDmS und mDoS.

Der CFI reagiert noch stärker auf die Mehrfachevaluation durch Dozenten (vgl. Modell ML-CFA in Abbildung 9). Während der CFI für das Modell oDoS mit .956 noch akzeptabel ist, liegt er mit .896 für das Modell mDoS deutlich unter der Grenze für ein akzeptables Modell (siehe auch Tabelle 4). Messmodelle für Fragebögen zur LVE würde man in diesem Fall nach den Kriterien für einen guten RMSEA zwar beibehalten, nach den Kriterien für einen guten CFI jedoch verwerfen. Die Mehrfachevaluation auf Dozentenebene hat demnach einen Einfluss auf die Beurteilung der Modellgüte. Im Vergleich dazu lassen beide Fitmaße eine positive Interpretation bzgl. der Modellgüte zu, wenn eine Stichprobe ohne Mehrfachevaluation

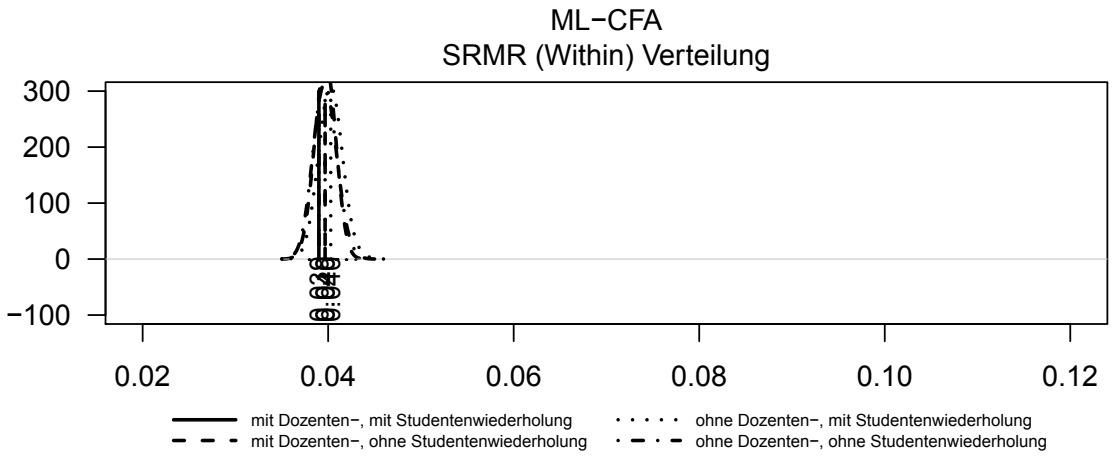


Abbildung 10. Verteilung des SRMR-Within für die ML-CFA zur Beurteilung des Modell-Fits in Abhängigkeit des Stichprobentyps.

vorliegt.

Durch die Betrachtung der globalen Fitmaße RMSEA und CFI kann eine ebenenspezifische Missspezifikation des Modells nicht identifiziert werden. Vor allem einen Fehlspezifikation auf dem Between-Level wird von den globalen Fitmaßen nicht erkannt. Aus diesem Grund werden hier auch für Within- und Between-Level spezifische Fit-Maße des SRMR angegeben. Für das Within-Messmodell sind die SRMR-Werte aller Modelle kleiner als die erforderliche kritische Größe von .05, die auf einen guten Modellfit hindeutet (vgl. Abbildung 10). Bei der Betrachtung des $\text{SRMR}_{between}$ fällt auf, dass dieser größer als der SRMR_{within} ist. Dennoch ist er deutlich kleiner als .10 und damit als akzeptabel zu interpretieren (vgl. Abbildung 11). Für beide SRMR-Maße zeigt sich ein deutlich geringer Einfluss der Mehrfachevaluation auf den entsprechenden SRMR-Kennwert als für den RMSEA und den CFI (vgl. Abbildung 10 und Abbildung 11).

Faktorwerte. Die Korrelationen der Faktorwerte wurden zunächst für jede Dimension des PELVE separat berechnet und anschließend gemittelt. Die Tabelle 5 zeigt die gemittelten Korrelationen der Faktorwerte zwischen verschiedenen Bedingungen der Mehrfachevaluation und in Abhängigkeit des CFA-Verfahrens. Bei der CFA auf Veranstaltungsebene (vgl. CFA-V in Tabelle 5) unterscheiden sich die Korrelationen der Faktorwerte in Abhängigkeit der Mehrfachevaluation. Die Faktorwerte aus der Gesamtstichprobe (mDmS) korrelieren nur zu .93 mit Faktorwerten, die auf Basis der Stichproben ohne Mehrfachevaluation

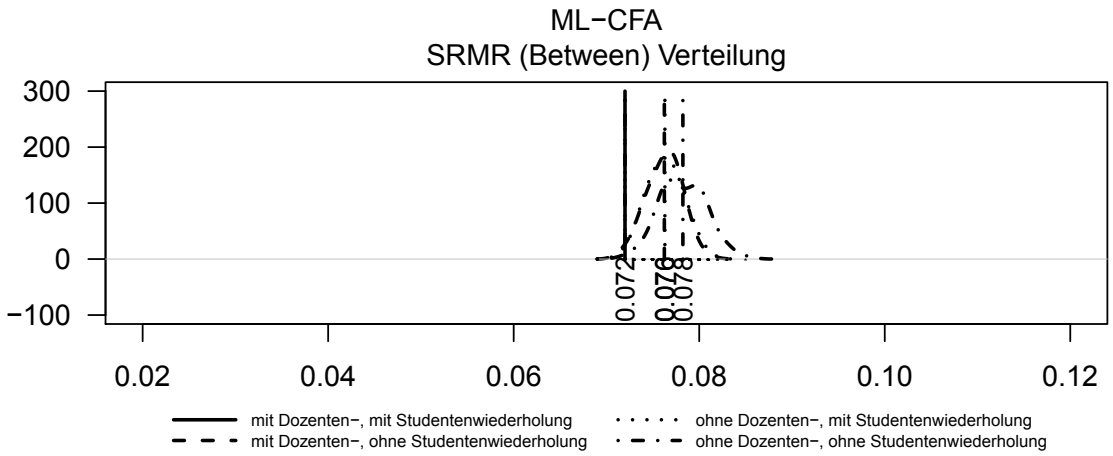


Abbildung 11. Verteilung des SRMR-Between für die ML-CFA zur Beurteilung des Modell-Fits in Abhängigkeit des Stichprobentyps.

(oDoS) geschätzt wurden. Ein sehr ähnliches Muster zeigt sich für die CFA auf Studentenebene (vgl. CFA-S in Tabelle 5). Auch hier sind die Korrelationen zwischen oDoS und mDmS bzw. oDmS am niedrigsten. Wenn Mehrfachevaluation auf Studentenebene in den Stichproben enthalten ist, korrelieren die Faktorwerte geringer mit Faktorwerten der Stichprobe ohne Mehrfachevaluation (oDoS), als wenn dies nicht der Fall ist (vgl. oDoS und mDoS in Tabelle 5). Für die ML-CFA kann das Korrelationsmuster noch differenzierter betrachtet werden. Zunächst ist die sehr niedrige Korrelation von .88 zwischen der Gesamtstichprobe (mDmS) und Faktorwerten aus der Stichprobe ohne Mehrfachevaluation (oDoS) auffällig. Auch alle anderen Korrelationen sind niedriger als vergleichbare Korrelationen der CFA-Verfahren CFA-S und CFA-V. Tabelle 5 zeigt ebenfalls die Korrelationen der Faktorwerte, die durch unterschiedliche CFA-Verfahren geschätzt wurden, getrennt für jeden Stichprobentyp. Auffällig sind die vergleichsweise geringen Korrelationen zwischen der ML-CFA und den CFA-Verfahren CFA-S bzw. CFA-V. Zwischen den Verfahren CFA-S und CFA-V sind hingegen die Korrelationen der Faktorwerte höher.

Die Stabilität der Rangreihe von Faktorwerten ist für die CFA auf Studenten- bzw. Veranstaltungsebene demnach höher als für ML-CFA. Bei der ML-CFA ändert sich die Rangreihe stärker, wenn die Faktorwerte auf Basis unterschiedlicher Stichprobentypen geschätzt werden. Über die Verfahren hinweg ist die Korrelation der Faktorwerte auf Basis der oDmS-Stichprobe mit der Gesamtstichprobe (mDmS) am höchsten. Faktorwerte auf

Basis des Stichprobentyps oDoS korrelieren am geringsten mit den Faktorwerten auf Basis der Gesamtstichprobe (mDmS). Diese Verhältnisse sind auch bei der Betrachtung der Prozentrangdifferenzen zu beobachten. Eine grafische Repräsentation der Verteilung von Prozentrangdifferenzen zeigt Abbildung 12. Hier fallen die breiten Verteilungen der Prozentrangdifferenzen beim Vergleich der CFA auf Studenten- bzw. Veranstaltungsebene mit der ML-CFA auf. Dieses Muster ist in Abbildung 12 für alle Stichprobentypen und Dimensionen des PELVE erkennbar. Die getrennte Darstellung für unterschiedliche latente Variablen in Abbildung 12 verdeutlicht zudem die Unterschiedlichkeit der Prozentrangdifferenzen in Abhängigkeit der jeweils betrachteten Dimension des PELVE.

Tabelle 5
Korrelationen der Faktorwerte in Abhängigkeit des verwendeten CFA-Verfahrens und Art der Mehrfachevaluation

		CFA-S				CFA-V				ML-CFA			
		mDmS	mDoS	oDmS	oDoS	mDmS	mDoS	oDmS	oDoS	mDmS	mDoS	oDmS	oDoS
CFA-S	mDmS	1											
	mDoS	.96	1										
	oDmS	1	.96	1									
	oDoS	.93	.98	.93	1								
CFA-V	mDmS	.95				1							
	mDoS		.96			.95	1						
	oDmS			.96		1	.96	1					
	oDoS				.96	.93	.98	.93	1				
ML-CFA	mDmS	.93				.94				1			
	mDoS		.91				.92			.90	1		
	oDmS			.96				.96		.97	.94	1	
	oDoS				.90				.91	.88	.97	.92	1

Anmerkungen: Dargestellt sind die Korrelationen der Faktorwerte zwischen unterschiedlichen CFA-Verfahren für die Stichprobenversionen mit Dozenten- und Studentenwiederholung (mDmS), ohne Dozenten- und Studentenwiederholung (oDoS), mit Dozentenwiederholung (mDoS) und mit Studentenwiederholung (oDmS).

Diskussion

In der vorliegenden Studie wurde der Einfluss der Mehrfachevaluation auf die Überprüfung des Messmodells eines Fragebogens zur LVE untersucht. Werden LVE-Ergebnisse

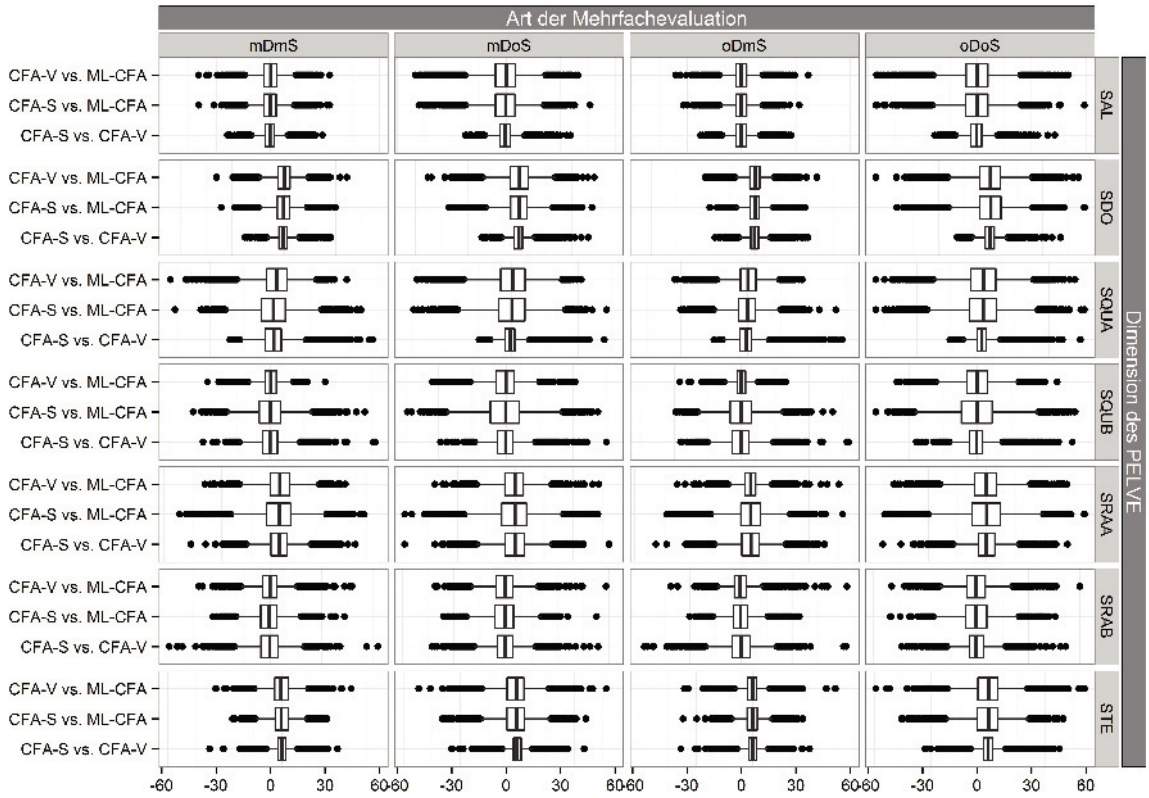


Abbildung 12. Prozentrangdifferenzen für die Dimensionen des PELVE in Abhängigkeit der Mehrfachevaluation und des CFA-Verfahrens.

über ihre primär beabsichtigte Feedbackfunktion hinaus verwendet, kann es durch die Zusammenführung veranstaltungsspezifischer Stichproben zu Gesamtstichproben zu Datenstrukturen kommen, deren Bedeutung in bisherigen Untersuchungen zur Modellpassung der eingesetzten Fragebögen nur wenig untersucht wurde. Die Hochschulforschung widmet sich jedoch verstärkt Fragestellungen zur Entwicklung der Lehre und zu Potentialen, diese gezielt zu verbessern. Bei diesen weiterführenden Fragestellung über die Feedbackfunktion hinaus, werden oftmals vorhandene veranstaltungsspezifische Stichproben zusammengeführt, ohne die Mehrfachevaluation durch Studenten und Dozenten zu beachten. Für die Untersuchung der Entwicklung von LVE-Ergebnissen über die Zeit oder für die Evaluation hochschuldidaktischer Maßnahmen, ist die Frage nach der Dimensionalität von ebenso großer Bedeutung wie bei der nachträglichen Aggregation der LVE-Ergebnisse für die Verwendung auf Steuerungsebene. Zur Überprüfung der Dimensionalität wird in der LVE hauptsächlich die CFA auf Veranstaltungsebene oder Studentenebene eingesetzt (vgl. Rindermann, 2009; Schmidt & Loßnitzer, 2010; Spooren et al., 2013). Die Ergebnisse der Studie

zeigen, dass selbst nach vorangehender Aggregation der Studentenurteile zu veranstaltungsspezifischen Itemmittelwerten bei der CFA auf Veranstaltungsebene eine Verschlechterung des Modellfits zu beobachten ist, wenn Studierende an mehreren Evaluationen teilgenommen haben. Für mehrere Veranstaltungen desselben Dozenten in der Gesamtstichprobe ist dieser negative Effekt auf die Kennwerte des Modellfits nicht zu beobachten. Entgegengesetzt verhält es sich bei der ML-CFA. Hier wird das Design der LVE durch ein separates Modell innerhalb der Veranstaltungen (auf Studentenebene) und zwischen Veranstaltungen (Veranstaltungsebene) am besten abgebildet. Mehrfach evaluierende Studenten in der Stichprobe führen nicht zu einer Verschlechterung des Modellfits. Mehrfachevaluation auf Dozentenebene hingegen verschlechtert den Modellfit, weil eine zusätzlich relevante Hierarchieebene in der Stichprobe vorhanden ist, die bei der Analyse nicht berücksichtigt wird. Bei der CFA auf Studentenebene sind keinerlei Effekte der Mehrfachevaluation auf die Kennwerte des Modellfits auszumachen. Problematisch ist jedoch, dass die so konstruierten latenten Variablen nur auf Studentenebene, nicht aber auf Veranstaltungsebene definiert sind, sodass eine Modellierung der LVQ für weiterführende Fragestellungen nur schwer möglich ist. Zudem bildet das Modell damit nur unzureichend das theoretische Messmodell von Fragebögen zu LVE ab, das sich auf Dimensionen der Veranstaltungsqualität bezieht und hierfür belastbare Aussagen ermöglichen möchte.

Die vorliegende Studie verdeutlicht damit, dass die Berücksichtigung der Mehrfachevaluation auf Studentenebene nicht notwendig ist, sofern eine ML-CFA zur Prüfung des Messmodells angewendet wird. Die Mehrfachevaluation auf Dozentenebene ist für die Prüfung des Messmodells hingegen auszuschließen oder adäquat zu berücksichtigen. Einschränkend sei an dieser Stelle erwähnt, dass der vorliegenden Studie ein Fragebogen zugrunde liegt, der explizit Facetten der Lehrveranstaltungsqualität auf Veranstaltungsebene postuliert. Handelt es sich bei einzelnen Dimensionen und den zugehörigen Items um eine Selbsteinschätzung der Studenten, sodass veranstaltungsspezifische Aspekte nur indirekt auf die Beurteilung wirken, ist erneut zu prüfen, ob ein Multilevel-Messmodell adäquat ist. Dies ist dann der Fall, wenn die Ergebnisse der LVE durch veranstaltungsspezifische Eigenschaften beeinflusst werden und latente Variablen auf Veranstaltungsebene der Theorie

nach zu konstruieren sind.

Bezüglich steuerungsrelevanter Informationen versucht die Studie, mit Hilfe der Betrachtung von Korrelationen und Prozentrangdifferenzen, den Einfluss der Mehrfachevaluation auf den Rangplatz einer Veranstaltung zu untersuchen. Einen Einfluss der Mehrfachevaluation auf die Rangplätze der Veranstaltungen war zu erwarten, weil sich die Antwortverteilungen und damit auch Mittelwerte und Faktorwerte der Veranstaltungen zwischen den unterschiedlichen Stichprobentypen unterscheiden. Diese Unterschiede sind zwar durch die zufällige Auswahl nicht signifikant, jedoch ausreichend, um eine Rangplatzveränderung zu beobachten. Der Einfluss der Mehrfachevaluation kann demnach nicht allein auf die Mehrfachevaluation zurückgeführt, sondern nur in Relation zueinander interpretiert werden. Die Ergebnisse zeigen eine stärkere Veränderung innerhalb der ML-CFA als bei den anderen Verfahren. Dies liegt möglicherweise an der zusätzlichen Messfehlerbereinigung auf Veranstaltungsebene im Rahmen der ML-CFA. Hier werden die Intercepts (Veranstaltungsmittelwerte) als messfehlerbehaftete Größen im Modell verwendet. Der Einfluss der Mehrfachevaluation kann aus diesem Grund nicht handlungsleitend für die Wahl des korrekten Stichprobentyps sein. Vielmehr eignet sich aus theoretischen Gründen der Stichprobentyp ohne Dozenten- und mit Studentenwiederholung als Grundlage für die Schätzung der Veranstaltungsqualität auf den verschiedenen Dimensionen. Dieser Stichprobentyp nutzt alle Informationen auf Studentenebene und damit die vollständige Veranstaltungsevaluation wie sie vom LVE-Design vorgesehen ist und vermeidet es, dass Dozenten mit mehreren Veranstaltungen im Vergleich zu sich selbst stehen, wenn Rangplätze betrachtet werden. Gleichzeitig sind die Ergebnisse der ML-CFA für diesen Stichprobentyp zufriedenstellend. Die Prozentränge zeigen dennoch einen deutlichen Einfluss der CFA-Verfahren auf die Rangreihe der Veranstaltung bei gleichem Stichprobentyp. Von Entscheidungen auf Basis einer derart errechneten Rangreihe von Veranstaltungen ist daher abzuraten.

Literatur

Asparouhov, T. & Muthén, B. (2010). *Plausible values for latent variables using Mplus.* Zugriff auf `http://www.statmodel.com/\-download/\-Plausible.pdf`

Asparouhov, T. & Muthén, B. O. (2006). *Comparison of estimation methods for complex survey data analysis*. Unpublished manuscript, UCLA.

Browne, M. W. & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Hrsg.), *Testing structural equation models* (S. 136–162). Thousand Oaks, CA: Sage.

Chung, H. & Beretvas, S. N. (2012). The impact of ignoring multiple membership data structures in multilevel models. *British Journal of Mathematical and Statistical Psychology*, *65* (2), 185–200. doi: 10.1111/j.2044-8317.2011.02023.x

Clayson, D. E. (2007). Conceptual and statistical problems of using between-class data in educational research. *Journal of Marketing Education*, *29*, 34–38.

Fielding, A. & Goldstein, H. (2006). *Cross-classified and multiple membership structures in multilevel models: An introduction and review* (Research Report Nr. 791). Department for Education and Skills.

Hu, L. & Bentler, P. (1995). Evaluating model fit. In R. H. Hoyle (Hrsg.), *Structural equation modeling. Concepts, issues, and applications* (S. 76–99). London: SAGE.

Hu, L. & Bentler, P. M. (1998). Fit indices in covariance structure analysis: Sensitivity to under-parameterized model misspecification. *Psychological Methods*, *3* (4), 424–453. doi: 10.1037/1082-989X.3.4.424

Loßnitzer, T., Schmidt, B. & Born, S. (2007). Zentrale Lehrveranstaltungsevaluation an der Friedrich-Schiller-Universität Jena - Qualitätsmodell und Messinstrument. In M. Krämer, S. Preiser & K. Brusdeylins (Hrsg.), *Psychologiedidaktik und Evaluation VI*. (S. 327–335). Göttingen: V&R unipress.

Marsh, H. W. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. *Journal of Educational Psychology*, *75* (1), 150–166.

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliablity, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (Hrsg.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (S. 319–383). Dordrecht: Springer.

Marsh, H. W., Muthén, B. O., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S. & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling*, *16*, 439–476.

Marsh, H. W. & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist*, *53*, 1187–1197.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika*, *49* (1), 115–132.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22* (3), 376–398. doi: 10.1177/0049124194022003006

Muthén, B. O. & Asparouhov, T. (in Druck). Item response modeling in Mplus: A multi-dimensional, multi-level, and multi-timepoint example. In W. J. van der Linden & R. K. Hambleton (Hrsg.), *Handbook of item response theory: Models, statistical tools, and applications*. Boca Raton, FL: Chapman & Hall/CRC Press. Zugriff auf `http://www.statmodel.com/download/IRT1Version2.pdf`

Muthén, L. K. & Muthén, B. O. (2008). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

R Core Team. (2014). R: A language and environment for statistical computing [Software-Handbuch]. Wien, Österreich. Zugriff auf `http://www.R-project.org/`

Rindermann, H. (2009). *Lehrevaluation: Einführung und Überblick zu Forschung und Praxis der Lehrveranstaltungsevaluation an Hochschulen mit einem Beitrag zur Evaluation computerbasierten Unterrichts* (2. Aufl.). Landau: Empirische Pädagogik e. V.

Ryu, E. & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling*, *16*, 583–601. doi: 10.1080/10705510903203466

Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8* (2), 23–74.

Schmidt, B. & Loßnitzer, T. (2010). Lehrveranstaltungsevaluation: State of Art, ein Definitionsvorschlag und Entwicklungslinien. *Zeitschrift für Evaluation*, *9* (1), 49–72.

Sengewald, E. & Vetterlein, A. (2015). Multilevel Faktorenanalyse für Fragebögen zur Lehrveranstaltungsevaluation. *Diagnostica*, *61*, 116–123.

Skinner, C. J., Holt, D. & Smith, T. M. F. (Hrsg.). (1989). *Analysis of complex surveys*. West Sussex, England: Wiley.

Spooren, P., Brockx, B. & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research*, *83* (4), 598–642.

Toland, M. D. & de Ayala, R. J. (2005). A multilevel factor analysis of students' evaluations of teaching.

Educational and Psychological Measurement, *65* (2), 272–296.

Vetterlein, A. & Sengewald, E. (2015). Ergebnisdarstellung in der Lehrveranstaltungsevaluation. Effekte verschiedener Berichte auf die Qualität von Lehrveranstaltungen. *Diagnostica*, *61*, 153–162.

6 Manuskript 3

Diagnostica, 61 (3), 153 – 162 © Hogrefe Verlag, Göttingen 2015

Ergebnisdarstellung in der Lehrveranstaltungsevaluation

Effekte verschiedener Berichte auf die Qualität von Lehrveranstaltungen

Anja Vetterlein und Erik Sengewald

Zusammenfassung. Nach Marsh (2007) ist eine Funktion der Lehrveranstaltungsevaluation das diagnostische Feedback und dient der Verbesserung der Lehre. Doch der Weg vom Ergebnisbericht bis zur Veränderung der nächsten Lehrveranstaltung ist „weit und beschwerlich", wie Helmke und Hosenfeld (2005) in ihrem Rezeptionsmodell darlegen. Zusätzlich zur bereits etablierten langen Ergebnisdarstellung wird eine neue kompakte Ergebnisdarstellung für den Kontext der Lehrveranstaltungsevaluation entwickelt. Die Studie untersucht in einem randomisierten Experiment mit $N = 283$ Dozenten die Wirkung der beiden Ergebnisdarstellungen auf die Lehrveranstaltungsqualität. Letztere erfasst der Fragebogen PELVE (Born, Loßnitzer & Schmidt, 2006) auf sieben latenten Dimensionen. Es wird ein Multi-Level-Strukturgleichungsmodell für kategoriale Variablen in M*plus* spezifiziert. Die Ergebnisse zeigen, dass Dozenten mit dem kompakten Bericht höhere Werte auf der Bewertungsdimension Begleitmaterialien erreichen. Trotz höherer Komplexität finden sich keine Hinweise auf negative Effekte des kompakten Ergebnisberichts auf die Qualität der folgenden Lehrveranstaltung.
Schlüsselwörter: Hochschulforschung, Lehrveranstaltungsevaluation, Rückmeldeformat, Randomisiertes Experiment, Rezeptionsforschung

Result Reports for Students' Evaluations of Teaching: Effects of Different Reports on Course Quality

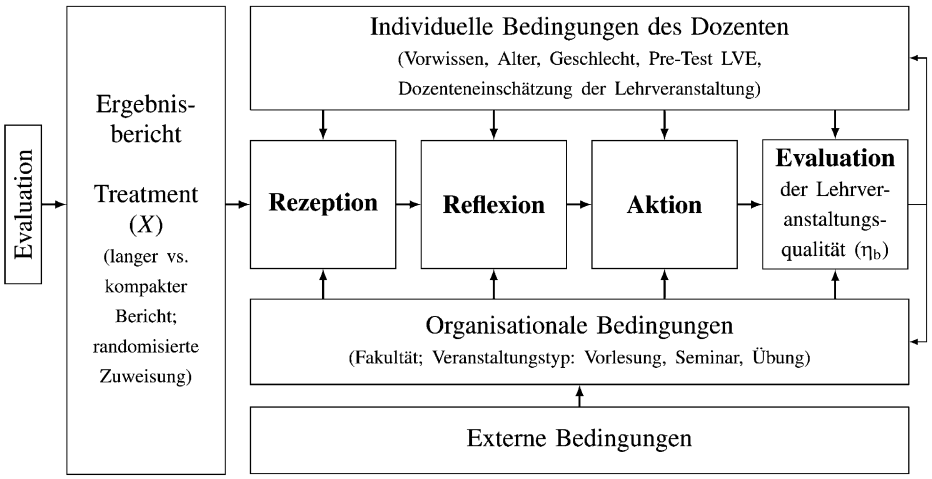
Abstract: According to Marsh (2007) students' evaluations of teaching (SET) are collected to provide diagnostic feedback to teachers for improving teaching. However, the way from evaluation to innovation is far and troublesome, as Helmke and Hosenfeld (2005) describe in their perception model. In addition to the already existing long report we developed a new compact report for SET data. We examine in a randomized experiment with $N = 283$ lecturers, the effect of the two reports on the course quality. Course quality is measured by the questionnaire PELVE (Born, Loßnitzer, & Schmidt, 2006) on seven latent dimensions. A multi level structural equation model for categorical variables is specified in M*plus*. The results show that lecturers with the compact report achieve higher scores on the dimension course material. Despite higher information density, we find no evidence for negative effects of the compact report on the course quality.
Keywords: higher education research, students' evaluations of university teaching, presentation of results, randomized experiment

Die *Lehrveranstaltungsevaluation* (LVE) dient an deutschen Hochschulen häufig als *ein* Baustein zur Qualitätssicherung in Lehre und Studium. An der Friedrich-Schiller-Universität Jena (FSU Jena) erfolgt die LVE sowohl dozentenorientiert und freiwillig (vgl. Evaluationsordnung der FSU Jena, 2012) als auch systematisch, fair, ökonomisch und wirksam (Steyer, Schmidt & Loßnitzer, 2004). Neben weiteren Zielen von LVE nennt Marsh (2007) allen voran das formative/diagnostische Feedback zur Verbesserung der eigenen Lehre. Ähnlich argumentiert Rindermann (2001, S. 79), für den LVE darauf abzielt, „der Lehrkraft selbst bei der Verbesserung der eigenen Didaktik und Veranstaltungsgestaltung [zu] helfen". Die aufgeführten Zielformulierungen von Marsh (2007), Rindermann (2001) und Steyer et al. (2004) machen deutlich, dass *hohe Wirksamkeitserwartungen* mit der Durchführung von LVE verbunden sind, die sich sowohl auf die Verbesserung einzelner Lehrveranstaltungen als auch auf das gesamte Qualitätsentwicklungssystem einer Hochschule (Schmidt & Loßnitzer, 2010) beziehen. Unbearbeitet ist bisher die Lücke zwischen den LVE-Ergebnissen eines Dozenten und der Ableitung von Veränderungsmaßnahmen. Müller (2010) weist darauf hin, dass der bloße Empfang von Ergebnissen nicht ausreicht, sondern die (statistischen) Ergebnisse auch *verstanden* werden müssen, um anschließend die *richtigen* Maßnahmen daraus ableiten zu können.

Im Rahmen der LVE wird häufig und ausgiebig über die Formulierung von Items diskutiert; über die Darstellung der erfassten Daten hingegen nahezu nie. Bislang

Dieses Vorhaben wurde aus Mitteln des Bundesministeriums für Bildung und Forschung (Förderkennzeichen 01PL12071) gefördert. Die Autoren danken zwei anonymen Gutachtern für wertvolle Hinweise zum Manuskript.

DOI: 10.1026/0012-1924/a000128



Anmerkung: Die vier Schritte des Rezeptionsprozesses sind fett gedruckt.

Abbildung 1. Rezeptionsmodell in Anlehnung an Helmke und Hosenfeld (2005) übertragen auf den Hochschulkontext.

erfolgt die Ergebnisdarstellung in der LVE intuitiv, da nicht mehr als „Empfehlungen" (Rindermann, 2009, S. 273) zur Gestaltung von Ergebnisrückmeldungen vorliegen. Dabei existieren viele unterschiedliche Möglichkeiten für die Ergebnisrückmeldung an deutschen Hochschulen (z. B. schriftlicher Bericht vs. mündliches Feedback (vgl. *Teaching Analysis Poll* von Frank, Fröhlich & Lahm, 2011); statistischen Kennzahlen vs. grafische Darstellung; Balkengrafiken vs. Verteilungsgrafiken vs. Profillinien).

Fiege (2013) empfiehlt daher, die Rückmeldeformate im Rahmen systematischer Rezeptionsforschung zu evaluieren. Die vorliegende Studie stellt (a) eine konventionelle und lange und (b) eine neue und kompakte Art der Darstellung von LVE-Ergebnissen vor und untersucht den Effekt der beiden Rückmeldeformate auf die Lehrveranstaltungsqualität.

Theorie

Lehrveranstaltungsqualität (LVQ)

Die Diskussion über die Definition und Messung von *guter Lehre* im Allgemeinen und der *Qualität von Lehrveranstaltungen* im Speziellen ist lang und umfangreich (für einen Überblick siehe Schmidt, 2007). Trotzdem herrscht inzwischen weitgehend Einigkeit darüber, dass die Qualität einer Lehrveranstaltung ein mehrdimensionales Konstrukt ist (Feldman, 1976; Marsh, 2007) und es deshalb auch mehrdimensional gemessen werden muss. Neben weiteren Fragebögen zur LVE (z. B. SEEQ, Marsh, 1982; FEVOR, Staufenbiel, 2000; TRIL, Gollwitzer & Schlotz, 2003) erfüllt auch der an der FSU Jena eingesetzte Fragebogen PELVE (*Prozess- und Ergebnisorientierte Lehrveranstaltungsevaluation*; Born, Loßnitzer & Schmidt, 2006) den Anspruch der multidimensionalen Messung von LVQ.

Rezeption von Ergebnisberichten

Eine solide Evaluation führt keineswegs zwangsläufig zu Verbesserungen, so Helmke und Hosenfeld (2005). Vielmehr beschreiben die Autoren in ihrem Rezeptionsmodell (vgl. Abbildung 1) den Weg von der Information/Evaluation (Bericht der LVE) zur Innovation (Verbesserungen in der Lehre) als „weit und beschwerlich" (Helmke & Hosenfeld, 2005, S. 147). Das Modell beschreibt den idealen Rezeptionsprozess in vier Schritten: (1) Rezeption: Wahrnehmung der Information, (2) Reflexion: Analyse der Information, (3) Aktion: Entwicklung und Umsetzung von Maßnahmen und (4) Evaluation: Überprüfung der Wirkung der Maßnahmen. Die einzelnen Schritte werden zudem von individuellen, externen und organisationalen Bedingungen beeinflusst. Der Rezeptionsprozess hat viele Unbekannte und ist an vielen Stellen vulnerabel. Es wird davon ausgegangen, dass die Schritte aufeinander aufbauen. Misslingt einer der Schritte, so scheitert der gesamte Rezeptions- und Verbesserungsprozess. Das Rezeptionsmodell verdeutlicht den großen (zeitlichen) Abstand, der zwischen dem Empfang der Ergebnisse, der Informationsverarbeitung und der folgenden Lehrveranstaltung (inklusive erneuter Evaluation) liegt.

In einer aktuellen Nutzerbefragung an der FSU Jena zum Umgang mit Lehrevaluationsberichten gaben alle der befragten Dozenten ($N = 279$) an, den Bericht durchgelesen zu haben. Für die vorliegende Studie nehmen wir daher an, dass ein Rezeptionsprozess bei den Dozenten initiiert wurde.

Ergebnisdarstellung von LVE-Daten

Eine Möglichkeit Einfluss auf den Rezeptionsprozess zu nehmen, ist die Variation der Ergebnisdarstellung. In der Regel werden die umfangreichen Evaluationsdaten dem Dozenten mittels eines Ergebnisberichts zurückgemeldet. Gemäß Rindermann (2009, S. 276 f.) stellen Grafiken in

Ergebnisberichten dafür „eine besonders effektive Rückmeldungsform dar“ und „Stärken und Schwächen lassen sich durch ein grafisches Feedback schnell erfassen“. In den letzten Jahrzehnten stieg die Zahl von statistischen Verfahren und Datenmodellierung rapide an. Das Gebiet der grafischen Darstellung von Daten wurde dabei vernachlässigt, stellen Gelman und Unwin (2012) fest. Gleichwohl bezeichnen sie Datengrafiken als ein einfaches und geeignetes Mittel für die deskriptive und explorative Datenanalyse. Auch Tufte (2011, S. 9) sieht in Datengrafiken „the most effective way to describe, explore, and summarize a set of numbers“.

Unter Berücksichtigung der Hinweise zur exzellenten Grafikerstellung (im Sinne von Gelman & Unwin, 2012; Tufte, 2011) und auf Grundlage der Ergebnisse von Vortests (mit $N = 25$ Lehrenden und $N = 135$ Studierenden) wurde von den Autoren dieses Artikels an der FSU Jena eine *kompakte* Darstellung von Lehrevaluationsdaten konzipiert. Dabei bilden die Datengrafiken das Kernelement. Seit dem Sommersemester 2012 sind damit zwei Berichtsversionen (lang vs. kompakt) parallel im Einsatz.

Langer Bericht. Die lange Berichtsversion (siehe Abbildung 2) wird seit dem Wintersemester 2003/04 eingesetzt und umfasst drei Teile, die separat und nacheinander dargestellt werden: Teil 1 enthält Tabellen mit statistischen Kennwerten, Teil 2 die Darstellung der Daten als Balkengrafiken und Teil 3 die Darstellung der Daten als gestapelte Antwortverteilung. Daraus entsteht ein 26-seitiger Bericht. In dieser Berichtsversion erhält der Dozent die Ergebnisse zu einem Item an drei verschiedenen Positionen im Bericht: Zunächst werden die *deskriptiven Kennwerte in einer Tabelle* dargestellt und auf der Folgeseite wird die grafische Repräsentation der Mittelwerte in Form von *Balkengrafiken* (und Balken für die Vergleichswerte) abgedruckt. Nach der tabellarischen und grafischen Darstellung der Kennwerte aller Items, folgt die Abbildung der *Antwortverteilung durch gestapelte Häufigkeiten* am Ende des Berichts. Mit der Balkendarstellung erhält der Mittelwert ein starkes Gewicht. Der Mittelwert M (arithmetisches Mittel) ist ein Kennwert für die zentrale Tendenz der Verteilung der Antworten auf einem Item (Lokation) und markiert somit einen festen Punktwert. Die Darstellung des Mittelwerts M als einen flächigen Balken scheint daher fragwürdig. Eid, Gollwitzer und Schmitt (2010) empfehlen die Verwendung von Balkendiagrammen nur für die Darstellung von relativen und absoluten Häufigkeiten.

Kompakter Bericht. Der kompakte Bericht vereint alle drei Teile des langen Berichts in *einer kompakten Darstellung.* Der Itemtext, die statistischen Kennwerte und deren grafische Repräsentationen werden in einer Abbildung kombiniert (vgl. Abbildung 3). Die neu entwickelte *Datengrafik* bildet das Kernelement des kompakten Berichts. Sie erlaubt es, die Antwortverteilung eines Items, den sich daraus ergebenden Mittelwert, die Vergleichsmittelwerte und den Dozentenwert gemeinsam in einem Schaubild darzustellen. Der Mittelwert ist als senkrechter Strich abgetragen. Zwei kleine Dreiecke bilden die beiden Vergleichsmittelwerte ab und zeigen auf der angegebenen Skala mit der Spitze auf den exakten Wert. Ein Kreis repräsentiert den Dozentenwert.

Beide Berichte basieren auf identischen Informationen, die jedoch unterschiedlich dargestellt werden. Durch die Verdichtung der vielfältigen Informationen in einer Tabellenzeile, ist der kompakte Bericht deutlich *komplexer* als der lange Bericht.

Vorteile des kompakten Berichts. Mit nur 15 Seiten ist der kompakte Bericht zum einen *ökonomischer* als der lange Bericht (vgl. Anforderung nach Steyer et al., 2004), zum anderen erhöht der geringere Seitenumfang die Wahrscheinlichkeit, dass der Bericht vollständig gelesen wird. Die Grafik im kompakten Bericht betont die Antwortverteilung stärker als den Mittelwert. Das ist aus zwei Gründen von Vorteil: (1) Bei schief verteilten Daten auf den Items ist der Mittelwert als Kennwert für die zentrale Tendenz weniger aussagekräftig. (2) Die Antwortverteilung veranschaulicht die Variabilität der Studentenurteile. Diese Unterschiedlichkeit in den studentischen Antworten ist eine wichtige diagnostische Information für den Dozenten, die er in *einer* Grafik den anderen Kennwerten gegenüberstellen kann.

Nachteil des kompakten Berichts. Die kompakte Darstellung birgt das Risiko, dass die Evaluationsergebnisse falsch verstanden werden. Aufgrund der hohen Informationsdichte sind Details leicht zu übersehen. Nachtigall und Kröhne (2006, S. 70) weisen darauf hin, dass dadurch schnell die Gefahr besteht, „dass Ergebnisse […] ungerechtfertigt interpretiert werden“. Falsch interpretierte Ergebnisse würden inadäquate Veränderungen in der Lehre nach sich ziehen.

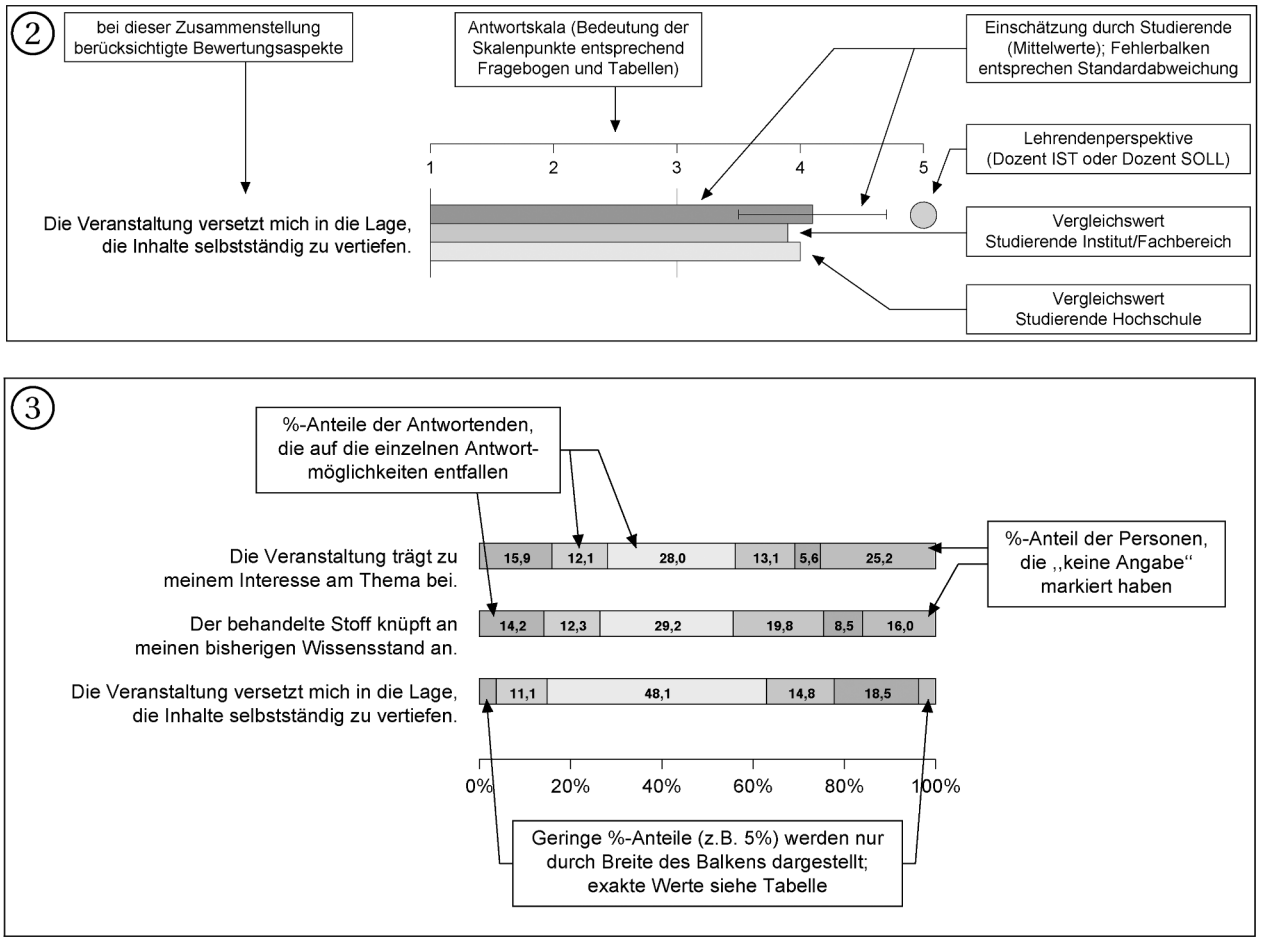
Fragestellung

LVE zielt darauf ab, den Dozenten ein Feedback zu ihren Lehrveranstaltungen zur Verfügung zu stellen. Damit verbunden ist die Hoffnung, einen Anstoß für Verbesserungen zu liefern. Die Rückmeldung an den Dozenten erfolgt in der Regel mit einem Ergebnisbericht, der im Prozess der LVE als Ausgangspunkt für Veränderungen dient.

Mit dem Empfang des Ergebnisberichts setzt der lange und umfangreiche Rezeptionsprozess (vgl. Abbildung 1) ein. Diese Untersuchung fokussiert auf den *Input* (*Ergebnisdarstellung*) und das *Outcome* (*Lehrveranstaltungsqualität*) des Rezeptionsprozesses. Zu diesem Zweck fand eine systematische Variation der Ergebnisdarstellung statt: Zusätzlich zur bereits bestehenden Ergebnisdarstellung

① **Aspekt** (*1 = stimme nicht zu* bis *5 = stimme zu*)

	Dozent IST	Veranstaltung					Hochschule	
		n	***M***	*SD*	*Min*	*Max*	***M***	*%*
Insgesamt gesehen, bin ich mit dieser Lehrveranstaltung zufrieden.	4	35	**3,3**	1,1	1	5	**4,1**	6,5
Insgesamt gesehen, bin ich mit den in dieser Veranstaltung erworbenen Qualifikationen zufrieden.	-	42	**3,1**	0,8	1	5	**3,8**	5,4



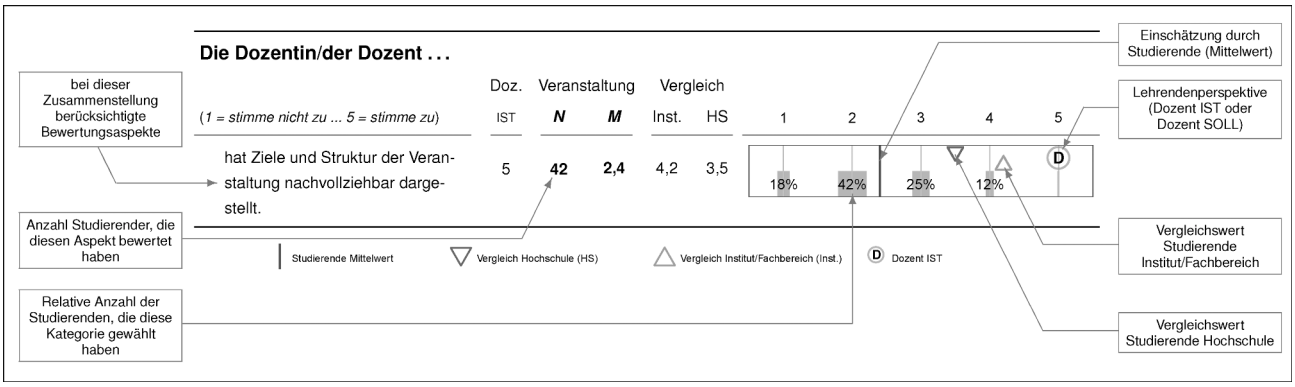
Anmerkungen: Alle drei Teile werden unabhängig und an unterschiedlichen Stellen im Ergebnisbericht abgedruckt: (1) Tabellen mit statistischen Kennwerten: absolute Häufigkeit der Antworten (*n*), Mittelwert (*M*), Standardabweichung (*SD*), Minimum (*Min*), Maximum (*Max*), Prozentrang (%); (2) Darstellung der Daten als Balkengrafiken; (3) Darstellung der Daten als gestapelte Antwortverteilung. Hervorhebungen und Formatierungen in der Abbildung entsprechen dem Original-Stimulusmaterial.

Abbildung 2. Beispiel für die dreiteilige Darstellung im langen Bericht.

(langer Bericht) wurde eine neue Darstellungsform entwickelt (kompakter Bericht).

Bislang fehlen empirische Studien zu Effekten von unterschiedlichen Rückmeldeformaten im Kontext der LVE. Daher sind die Auswirkungen einer kompakten Darstellung auf die Lehrveranstaltungsqualität im Vergleich zu einer langen Darstellung nicht bekannt. Aufgrund der oben genannten Vor- und Nachteile des kompakten Berichts sind sowohl positive als auch negative Effekte auf die LVQ denkbar.

Die Studie untersucht den Effekt der unterschiedlichen Ergebnisdarstellungen (lang vs. kompakt) auf die LVQ. Daraus ergibt sich die Frage: Welche Konsequenzen hat die Erhöhung der Darstellungskomplexität der Ergebnisberichte für die Lehrveranstaltungsqualität?



Anmerkungen: Die Darstellung enthält die folgenden Kennwerte: absolute Häufigkeit der Antworten (*N*), Mittelwert der Studenteneinschätzungen (*M*), Vergleichsmittelwert über alle Veranstaltungen des gleichen Instituts (Inst.), Vergleichsmittelwert über alle Veranstaltungen der gleichen Hochschule (HS). Hervorhebungen und Formatierungen in der Abbildung entsprechen dem Original-Stimulusmaterial.

Abbildung 3. Beispiel für die Darstellung im kompakten Bericht.

Methode

Abhängige Variablen: Messung von Lehrveranstaltungsqualität (LVQ) mit dem Instrument PELVE (η_b)

Die LVQ wird in dieser Studie über die Evaluation von Lehrveranstaltungen mit dem Instrument PELVE operationalisiert. Das multidimensional angelegte Instrument existiert in drei veranstaltungsspezifischen Versionen: (a) Fragebogen für Vorlesungen, (b) Fragebogen für Seminare und (c) Fragebogen für Übungen. Zur Beantwortung der Items steht eine fünfstufige Skala mit den Antwortpolen 1 = *stimme nicht zu* bis 5 = *stimme zu* zur Verfügung. Die erworbenen Kompetenzen werden von 1 = *wenig* bis 5 = *viel* eingeschätzt. Für jedes Item wird die zusätzliche Antwortoption *keine Angabe (k. A.)* dargeboten.

Mit der Durchführung der Lehrveranstaltungsevaluation geht die Nestung von Studenten (Erhebungseinheit) in Veranstaltungen (Analyseeinheit) einher. Die Mehrebenenstruktur der Daten (Level 1 within: Studentenebene; Level 2 between: Veranstaltungsebene) wird im *Multi-Level-Messmodell* des Instruments PELVE berücksichtigt (Sengewald & Vetterlein, 2013). Dieses Multi-Level-Messmodell (siehe Abbildung 4) postuliert jeweils sieben latente Variablen auf der Studenten- und Veranstaltungsebene (Modell-Fit-Indizes: RMSEA = .03; CFI = .96; TLI = .95; $SRMR_{within}$ = .04; $SRMR_{between}$ = .08). Aus 33 veranstaltungsübergreifenden Items des Fragebogens werden die folgenden sieben latenten Variablen auf Veranstaltungsebene konstruiert: (1) *Gesamteindruck* (Ges_b), (2) *Fachkompetenz* (FKo_b), (3) *sonstige Kompetenzen* (SKo_b), (4) *Rahmenbedingungen* (RaB_b), (5) *Begleitmaterialien* (BMa_b), (6) *Dozentenverhalten* (Doz_b) und (7) *Studentenverhalten* (Stu_b). Mit dem Instrument PELVE schätzen Studenten die *Lehrveranstaltungsqualität* auf diesen sieben genannten Dimensionen ein. Diese latenten Variablen auf der Veranstaltungsebene werden als die abhängigen Variablen (Outcome) für die Analysen in dieser Studie genutzt. Sie sind in dem Vektor $\boldsymbol{\eta}_b$ enthalten.

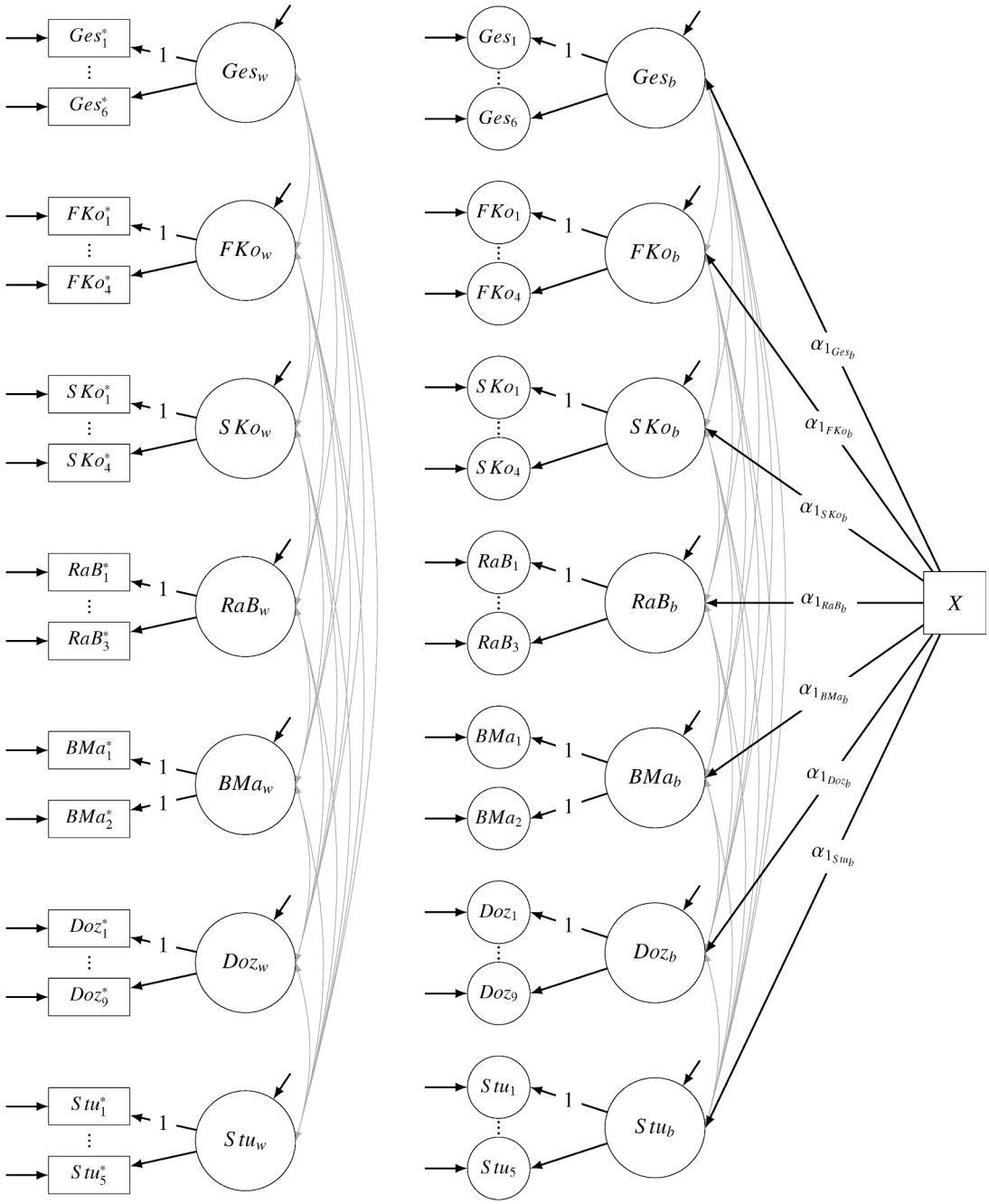
Unabhängige Variable *X*: Variation der Ergebnisdarstellung

Die Ergebnisdarstellung wird systematisch variiert (langer vs. kompakter Bericht) und randomisiert dem Dozenten zugewiesen. Die dichotome Treatmentvariable X (mit $X = 0$ für den langen Bericht und $X = 1$ für den kompakten Bericht) dient als unabhängige Indikatorvariable. Da das Treatment auf Veranstaltungsebene erfolgt, ist die Treatmentvariable X eine reine Level-2-Variable im Multi-Level-Messmodell (vgl. Abbildung 4). Die Zuteilung der Berichtsversion ist konstant für alle evaluierten Lehrveranstaltungen eines Dozenten über alle Semester.

Die Messung der abhängigen Variablen findet nach dem Treatment im Folgesemester statt: In einem Semester S hält ein Dozent eine Lehrveranstaltung und lässt diese evaluieren. Daraufhin erhält er per E-Mail einen langen oder einen kompakten Ergebnisbericht (Treatment). Die erste Evaluation im Folgesemester $S + 1$ wird für die Outcome-Messung herangezogen (vgl. Darstellung der Treatmentvariable X und Outcome-Variable $\boldsymbol{\eta}_b$ in Abbildung 1). Es handelt sich aufgrund dieses Designs um eine *ökologisch valide* Outcome-Messung, da sie in den standardisierten Ablauf der Lehrveranstaltungsevaluation nicht eingreift.

Experimentelles Design

Diese Untersuchung nutzt den „Gold Standard" (Rubin, 2008, S. 808) für die Analyse kausaler Effekte: das *randomisierte Experiment.* In vorangegangenen Pilotstudien



(a) Level 1: Studentenebene (b) Level 2: Veranstaltungsebene

Anmerkungen: Sieben Dimensionen des PELVE: Gesamteindruck (*Ges*), Fachkompetenz (*FKo*), sonstige Kompetenzen (*SKo*), Rahmenbedingungen (*RaB*), Begleitmaterialien (*BMa*), Dozentenverhalten (*Doz*) und Studentenverhalten (*Stu*). Index *w:* within (Level 1); Index *b:* between (Level 2). * Kennzeichnung von Latent Response-Variablen des jeweiligen Items.

Abbildung 4. Pfaddiagramm des Multi-Level-Messmodells des Instruments PELVE (Sengewald & Vetterlein, 2013) mit einer dichotomen Treatmentvariablen *X*.

haben sich Wissenstests und subjektive Selbsteinschätzungen zum Verständnis der Ergebnisdarstellungen im Rahmen von Online-Umfragen als unbrauchbar und zudem als nicht ökologisch valide erwiesen. Der fragile Rezeptionsprozess enthält viele Mediatoren und (z. T. unbekannte) Kovariaten, die schwer operationalisierbar und reliabel messbar sind. Ohne alle relevanten Kovariaten ist eine erwartungstreue Schätzung des kausalen Effekts im quasi-experimentellen Design nicht möglich (Steyer, Mayer & Fiege, 2014). Im randomisierten Expe-

riment hingegen soll mit Hilfe der zufälligen Zuweisung sichergestellt werden, dass die Wahrscheinlichkeit in die Treatmentgruppe zu kommen für jede Person gleich ist. Damit kann angenommen werden, dass es keine systematischen Unterschiede zwischen den Gruppen bezüglich relevanter Kovariaten vor dem Treatment gibt. Die Treatmentwahrscheinlichkeit ist demzufolge nicht abhängig von den Kovariaten der Person (z. B. *Geschlecht*, *Lehrerfahrung*, *Akademischer Grad*). Somit darf die Unverfälschtheit der Treatment-Regression $E(\boldsymbol{\eta_b} \mid X)$ angenommen werden (Steyer, Mayer & Fiege, 2014; vgl. Gleichung 1). In einem randomisierten Experiment dient die Erwartungswertdifferenz des Outcomes der beiden Gruppen (vgl. Gleichung 3) als Schätzer für den kausalen Effekt (Fiege, Kröhne & Steyer, 2010).

Datenanalyse

Aufgrund der oben beschriebenen Randomisierung ist die Zuordnung zu einer der Experimentalgruppen auf der theoretischen Ebene *nicht* abhängig von den Kovariaten (z. B. *Geschlecht*, *Akademischer Grad*, *Lehrveranstaltungstyp*). Dennoch kann es bei der empirischen Anwendung in der konkreten Stichprobe zufällig zu unterschiedlichen Verteilungen der Kovariaten in den Experimentalgruppen kommen. Daher wird mit χ^2-Tests die Unabhängigkeit der genannten nominalen Kovariaten von der Treatmentzuweisung für die Stichprobe überprüft.

Zur Effektschätzung wurde ein Multi-Level-Strukturgleichungsmodell (vgl. Abbildung 4) für *kategoriale* Variablen berechnet, da nicht von normalverteilten, intervallskalierten Daten ausgegangen werden konnte. Die Abstände zwischen den vier Schwellenparametern der fünfstufigen Ratingskala waren nicht gleich (vgl. Muthen, 1984). Betrachtet werden die *Regressionen* der latenten Variablen auf Level 2 auf die Treatmentvariable X (vgl. Gleichung 1; wobei $\boldsymbol{\eta_b}$ der Vektor der latenten Variablen, $\boldsymbol{\alpha_{0b}}$ der Vektor für die Y-Achsenabschnitte und $\boldsymbol{\alpha_{1b}}$ der Vektor für die Steigungskoeffizienten ist). Gleichung 4 enthält die ausführliche Darstellung der Vektoren. Der Index b symbolisiert den Parameter auf Level 2 (between). Die Steigungskoeffizienten $\boldsymbol{\alpha_{1b}}$ dieser Regressionen können als Erwartungswertdifferenzen auf den abhängigen Variablen zwischen Kontrollgruppen (KG) und Treatmentgruppe (TG) interpretiert werden (vgl. Gleichung 3). Ist ein Steigungskoeffizient α_{1b} signifikant verschieden von 0, so unterscheiden sich die KG und TG statistisch bedeutsam voneinander bezüglich der betrachteten abhängigen latenten Variable η_b (Signifikanzniveau: $\alpha = .05$). Ein positiver Steigungskoeffizient α_{1b} gibt an, dass der Erwartungswert in der TG höher ist als in der KG. Für jede der sieben latenten Variablen wird durch die Differenz der Erwartungswerte in Kontroll- und Treatmentgruppe jeweils der durchschnittliche totale Effekt (ATE) geschätzt (vgl. Gleichung 3).

$$E(\boldsymbol{\eta_b} \mid X) = \boldsymbol{\alpha_{0b}} + \boldsymbol{\alpha_{1b}} X \quad (1)$$

$$\boldsymbol{\alpha_{0b}} = E(\boldsymbol{\eta_b} \mid X = 0) \quad (2)$$

$$\boldsymbol{\alpha_{1b}} = E(\boldsymbol{\eta_b} \mid X = 1) - E(\boldsymbol{\eta_b} \mid X = 0) \quad (3)$$

$$\begin{pmatrix} \eta_{\mathrm{Ges_b}} \\ \eta_{\mathrm{FKo_b}} \\ \eta_{\mathrm{SKo_b}} \\ \eta_{\mathrm{RaB_b}} \\ \eta_{\mathrm{BMa_b}} \\ \eta_{\mathrm{Doz_b}} \\ \eta_{\mathrm{Stu_b}} \end{pmatrix} = \begin{pmatrix} \alpha_{0\mathrm{Ges_b}} \\ \alpha_{0\mathrm{FKo_b}} \\ \alpha_{0\mathrm{SKo_b}} \\ \alpha_{0\mathrm{RaB_b}} \\ \alpha_{0\mathrm{BMa_b}} \\ \alpha_{0\mathrm{Doz_b}} \\ \alpha_{0\mathrm{Stu_b}} \end{pmatrix} + \begin{pmatrix} \alpha_{1\mathrm{Ges_b}} \\ \alpha_{1\mathrm{FKo_b}} \\ \alpha_{1\mathrm{SKo_b}} \\ \alpha_{1\mathrm{RaB_b}} \\ \alpha_{1\mathrm{BMa_b}} \\ \alpha_{1\mathrm{Doz_b}} \\ \alpha_{1\mathrm{Stu_b}} \end{pmatrix} (X) + \begin{pmatrix} \zeta_{\mathrm{Ges_b}} \\ \zeta_{\mathrm{FKo_b}} \\ \zeta_{\mathrm{SKo_b}} \\ \zeta_{\mathrm{RaB_b}} \\ \zeta_{\mathrm{BMa_b}} \\ \zeta_{\mathrm{Doz_b}} \\ \zeta_{\mathrm{Stu_b}} \end{pmatrix} \quad (4)$$

Mit der Software M*plus* Version 7.11 (Muthén & Muthén, 1998–2012) erfolgte die Spezifikation des Modells mit sieben abhängigen latenten Variablen (vgl. Gleichung 4). Der Y-Achsenabschnitt $\boldsymbol{\alpha_{0b}}$ ist auf 0 fixiert (Gleichung 2). Aufgrund des ordinalen Skalenniveaus der Daten und der fehlenden uni- und multivariaten Normalverteilung, wird der WLSMV-Schätzer (Muthén, du Toit, Spisic, 1997) verwendet. Um auszuschließen, dass bereits vor dem Treatment Unterschiede zwischen den Experimentalgruppen bestanden, wurden neben den Posttest-Daten auch die Pretest-Daten (Daten der vorangegangenen LVE) der gleichen M*plus*-Analyse unterzogen.

Stichprobe

Die Datenerhebung erfolgte an der FSU Jena begleitend zum operativen Tagesgeschäft der Lehrveranstaltungsevaluation. Die Untersuchung basiert auf einem Datensatz mit $N = 6\,892$ Bewertungen von Studenten, die $N = 283$ Lehrveranstaltungen evaluiert haben. Diese Lehrveranstaltungen wurden von $N = 283$ unterschiedlichen Dozenten gehalten, die zufällig der Kontrollgruppe KG ($n = 129$ Dozenten mit langem Bericht) oder der Treatmentgruppe TG ($n = 154$ Dozenten mit kompaktem Bericht) zugeteilt wurden. Jeder Dozent geht mit einer Lehrveranstaltung in den Datensatz ein (Mehrfachevaluation eines Dozenten ist ausgeschlossen). Die Verteilungen der Variablen *Geschlecht*, *Akademischer Grad* und *Lehrveranstaltungstyp* stellen Abbildung 5 und, ergänzend, die Tabelle des Elektronischen Supplements 1 dar.

Die Daten wurden vom Sommersemester 2012 bis einschließlich Sommersemester 2013 mit dem Instrument PELVE als Online- oder Papierversion an allen Fakultäten der FSU Jena erhoben. Konsistent mit den Befunden

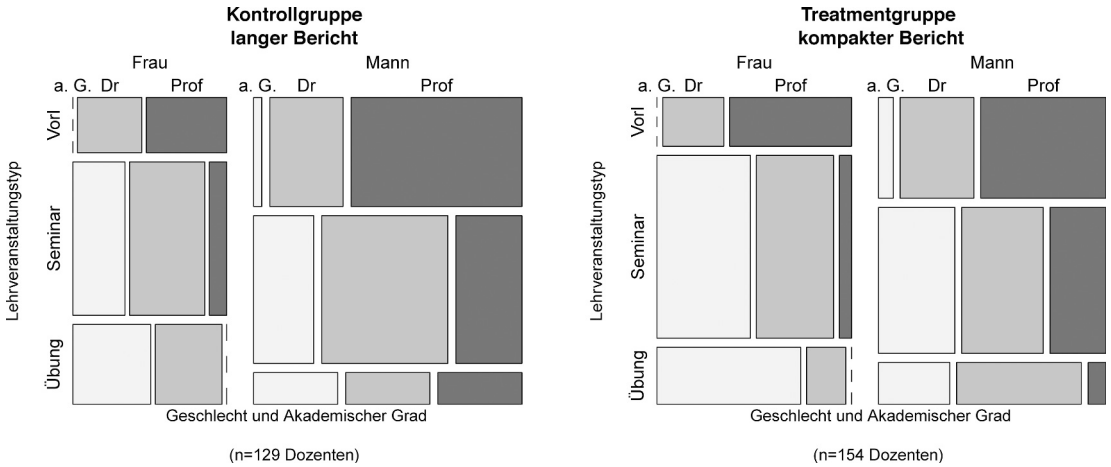


Abbildung 5. Deskriptiver Überblick über die Verteilung der Variablen Geschlecht (Frau – Mann), Akademischer Grad (anderer Grad (a. G.) – Doktor (Dr) – Professor (Prof)) und Lehrveranstaltungstyp (Vorlesung (Vorl) – Seminar – Übung) in der Kontrollgruppe und Treatmentgruppe.

Tabelle 1. Ergebnisse der M*plus*-Analysen zum Multi-Level-Modell mit latenten Regressionen auf die Treatmentvariable *X*

	Pretest-Vergleich von KG vs. TG				Posttest-Vergleich von KG vs. TG			
Dimensionen der Lehrveranstaltungsqualität	α_{1b}	***SE***	***p*-Wert**	***d***	α_{1b}	***SE***	***p*-Wert**	***d***
Gesamteindruck (Ges_b)	0.19	0.13	.46	0.19	0.10	0.13	.46	0.10
Fachkompetenz (FKo_b)	0.08	0.06	.21	0.17	0.12	0.07	.11	0.26
sonstige Kompetenzen (SKo_b)	0.13	0.08	.09	0.24	0.13	0.10	.18	0.24
Rahmenbedingungen (RaB_b)	0.09	0.08	.24	0.34	0.06	0.06	.30	0.24
Begleitmaterialien (BMa_b)	0.05	0.10	.62	0.06	0.39	0.15	.01*	0.51
Dozentenverhalten (Doz_b)	0.18	0.12	.14	0.18	0.15	0.13	.25	0.15
Studentenverhalten (Stu_b)	-0.01	0.08	.94	-0.01	0.08	0.08	.33	0.14

Anmerkungen: Vergleich von Kontrollgruppe (KG) und Treatmentgruppe (TG) auf den latenten Variablen, die die sieben Dimensionen der Lehrveranstaltungsqualität (LVQ) abbilden. SE = standard error. Signifikanzniveau: $\alpha = .05$; $^*p \leq .05$. Cohens *d* als Maß für die Effektstärke.

anderer Untersuchungen zur LVE (Daniel, 1996; Wolbring, 2013) ist auch bei dem Instrument PELVE die empirische Verteilung der Variablen schief und die Mittelwerte liegen auf einem hohen Niveau. Die durchschnittlichen Veranstaltungsmittelwerte der Dimensionen liegen zwischen $M_{SKo} = 3.26$ ($SD_{SKo} = 0.55$) und $M_{RaB} = 4.41$ ($SD_{RaB} = 0.32$).

Ergebnisse

Die KG und TG unterschieden sich nicht hinsichtlich der Variablen *Geschlecht* ($\chi^2_{(.05;\ 1,\ N=\ 283)} = 2.32$), *Akademischer Grad* ($\chi^2_{(.05;\ 2,\ N=\ 283)} = 5.37$) und *Lehrveranstaltungstyp* ($\chi^2_{(.05;\ 2,\ N=\ 283)} = 0.74$). Die entsprechenden χ^2-Tests sind nicht signifikant. Im Pretest bestehen keine signifikanten Mittelwertunterschiede zwischen KG und TG auf den sieben Dimensionen des PELVE (vgl. Tabelle 1).

Es wurde ein Multi-Level-Modell mit latenten Regressionen auf die Treatmentvariable *X* spezifiziert (Modell-Fit-Indizes: RMSEA = .02; CFI = .97; TLI = .97; $SRMR_{within} = .04$; $SRMR_{between} = .07$; $\chi^2 = 3\,087.68$; $df = 974$; *p*-Wert = .00). Entsprechend der Konventionen für deskriptive Fitmaße (Schermelleh-Engel, Moosbrugger & Müller, 2003), weist das Modell eine gute Passung auf. Die Ergebnisse der latenten Regressionen für den Posttest-Vergleich zwischen KG und TG sind in Tabelle 1 zusammengefasst.

Auf der latenten Variable *Begleitmaterialien* (BMa_b) unterscheiden sich die beiden Treatmentgruppen: Dozenten mit einem kompakten Bericht erreichen im Mittel höhere Werte auf der Dimension *Begleitmaterialien* (BMa_b) als Dozenten mit einem langen Bericht. Die Effektstärke liegt bei $d_{BMab =} 0.51$ und entspricht nach Cohen (1988) einem mittleren Effekt.

Auf den weiteren sechs latenten Variablen zeigen sich ebenfalls deskriptiv positive Regressionskoeffizienten α_{1b}; diese Unterschiede zwischen der Treatment- und der Kontrollgruppe sind jedoch nicht signifikant. Trotz

höherer Komplexität finden sich *keine* Hinweise auf *negative Effekte* des kompakten Ergebnisberichts auf die Qualität der folgenden Lehrveranstaltung. Ganz im Gegenteil: Es kann für den kompakten Ergebnisbericht von einem positiven Effekt auf einer Dimension der LVQ (BMa_b) berichtet werden.

Diskussion

Gegenstand der Studie war die Untersuchung der Effekte von unterschiedlichen Darstellungen der Evaluationsergebnisse auf die Lehrveranstaltungsqualität. Es zeigte sich ein positiver Effekt auf der Dimension *Begleitmaterialien* für Dozenten mit einem kompakten Bericht. Der positive Effekt zum Vorteil des kompakten Berichts ist vor dem Hintergrund der hohen zeitlichen Distanz zwischen Treatment und Outcome-Messung sowie dem relativ schwachen Treatment bemerkenswert.

Über den Rezeptionsprozess selbst ist weiterhin noch wenig bekannt, wenngleich durch die Ergebnisse erste Rückschlüsse auf den Prozess möglich sind: In der TG mit kompaktem Bericht war der Rezeptionsprozess erfolgreicher als in der KG. Weitere Studien sind erforderlich, um die Rezeption im Rahmen der LVE besser zu verstehen. Im Rahmen von Begleitforschung parallel zur operativen LVE lassen sich weitere relevante Kovariaten erheben (z. B. empfundene Länge des Berichts, subjektives Verständnis, Komplexität, Intensität der Rezeption etc.)

Neben der hier durchgeführten Schätzung von durchschnittlichen totalen Effekten (ATE) können auch *bedingte Effekte* geschätzt werden (Steyer, Mayer & Fiege, 2014). Damit lassen sich Fragen zur differentiellen Indikation der Ergebnisdarstellungen beantworten (z. B. Profitieren Naturwissenschaftler mehr von dem kompaktem Bericht als Geisteswissenschaftler?).

Wenn weitere Messzeitpunkte einbezogen werden, lässt sich unter anderem mit Hilfe der *Latent-State-Trait-Theorie* (Steyer, Mayer, Geiser & Cole, 2015) die Entwicklung von LVQ über die Zeit analysieren. Dabei werden sowohl Veranstaltungsspezifika (State) als auch Dozenteneigenschaften (Trait) in einer Analyse berücksichtigt.

Mit dieser Studie liegen empirische Daten zu den Effekten von unterschiedlichen Ergebnisdarstellungen vor. Die von den Autoren antizipierten negativen Effekte (aufgrund höherer Komplexität) konnten empirisch nicht bestätigt werden. Es wird daher empfohlen den kompakten Bericht in der LVE einzusetzen. Es konnte gezeigt werden, dass auch kleine Veränderungen im Rückmeldeformat Auswirkungen auf die LVQ haben können. Daher sollten ähnliche Entwicklungsmaßnahmen nicht ohne empirische Begleitung stattfinden und einer systematischen Evaluation unterzogen werden (vgl. Fiege, 2013).

Elektronische Supplemente (ESM)

Die elektronischen Supplemente sind mit der Online-Version diesess Artikels verfügbar über http://dx.doi.org/10.1026/0012-1924/a000128

ESM 1. Tabelle 2

Literatur

Born, S., Loßnitzer, T. & Schmidt, B. (2006). Lehrveranstaltungsevaluation an der Friedrich-Schiller-Universität Jena – Eine Analyse der Dimensionalität der eingesetzten Fragebögen. In B. Krause & P. Metzler (Hrsg.), *Empirische Evaluationsmethoden* (Bd. 10, S. 99 – 116). Berlin: ZeE.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* Hillsdale, NJ: Erlbaum.

Daniel, H.-D. (1996). Evaluierung der universitären Lehre durch Studenten und Absolventen. *Zeitschrift für Sozialisationsforschung und Erziehungssozialisation*, *16*, 149 – 164.

Eid, M., Gollwitzer, M. & Schmitt, M. (2010). *Statistik und Forschungsmethoden.* Weinheim: Beltz.

Feldman, K. A. (1976). The superior college teacher from the students' view. *Research in Higher Education*, *5*, *243 – 288.*

Fiege, C. (2013). *Faire Vergleiche in der Schulleistungsforschung – Methodologische Grundlagen und Anwendung auf Vergleichsarbeiten.* Dissertation, Friedrich-Schiller-Universität Jena.

Fiege, C., Kröhne, U. & Steyer, R. (2010). Theorie und Analyse kausaler Effekte. In H. Holling & B. Schmitz (Hrsg.), *Handbuch Statistik, Methoden und Evaluation* (S. 487 – 495). Göttingen: Hogrefe.

Frank, A., Fröhlich, M. & Lahm, S. (2011). Zwischenauswertung im Semester: Lehrveranstaltungen gemeinsam verändern. *Zeitschrift für Hochschulentwicklung*, *6 (3+4)*, 310 – 318.

Friedrich-Schiller-Universität Jena. (2012). Evaluationsstandards und Instrumente der Qualitätsentwicklung im Bereich Studium und Lehre (Evaluationsordnung). *Verkündungsblatt der Friedrich-Schiller-Universität Jena* (8/2015), 252 – 255.

Gelman, A. & Unwin, A. (2012). Infovis and Statistical Graphics: Different Goals, Different Looks. *Journal of Computational and Graphical Statistics*, *22*, 2 – 28.

Gollwitzer, M. & Schlotz, W. (2003). Das Trierer Inventar zur Lehrveranstaltungsevaluation (TRIL): Entwicklung und erste testtheoretische Erprobungen. In G. Krampen & H. Zayer (Hrsg.), *Psychologiedidaktik und Evaluation IV* (S. 114 – 128). Bonn: Deutscher Psychologen Verlag.

Helmke, A. & Hosenfeld, I. (2005). Standardbezogene Unterrichtsevaluation. In G. Brägger, B. Bucher & N. Landwehr (Hrsg.), *Schlüsselfragen zur externen Schulevaluation* (S. 127 – 151). Bern: Hep-Verlag.

Marsh, H. W. (1982). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology*, *52*, 77 – 92.

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliablity, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in Higher Education: An*

evidence-based perspective (pp. 319 – 383). Dordrecht: Springer.

Müller, A. (2010). *Rückmeldungen nach Vergleichsarbeiten im Kontext des schulischen Qualitätsmanagements: Drei explorative Studien zu Gestaltung und Rezeption im Anschluss an KOALA-S.* Berlin: Mensch & Buch.

Muthén, B. O. (1984). A general Structural Equation Model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika*, *49*, 115 – 132.

Muthén, B. O., du Toit, S. H. C. & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes* [Bericht]. Zugriff am 16. 01. 2014 unter http://www.gseis.ucla.edu/faculty/muthen/articles/Article_075.pdf

Muthén, L. K. & Muthén, B. O. (1998 – 2012). *Mplus User's Guide.* (7th ed). Los Angeles, CA: Muthén & Muthén.

Nachtigall, C. & Kröhne, U. (2006). Methodische Anforderungen an schulische Leistungsmessung – auf dem Weg zu fairen Vergleichen. In H. Kuper & J. Schneewind (Hrsg.), *Rückmeldung und Rezeption von Forschungsergebnissen* (S. 59 – 74). Münster: Waxmann.

Rindermann, H. (2001). Die studentische Beurteilung von Lehrveranstaltungen – Forschungsstand und Implikationen. In C. Spiel (Hrsg.), *Evaluation universitärer Lehre – zwischen Qualitätsmanagement und Selbstzweck* (S. 61 – 88). Münster: Waxmann.

Rindermann, H. (2009). *Lehrevaluation: Einführung und Überblick zu Forschung und Praxis der Lehrveranstaltungsevaluation an Hochschulen* (2. Aufl.). Landau: Empirische Pädagogik e. V.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, *2*, 808 – 840.

Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the fit of Structural Equation Models: Tests of significance and descriptive Goodness-of-Fit Measures. *Methods of Psychological Research Online*, *8 (2)*, 23 – 74.

Schmidt, B. (2007). *Personalentwicklung für junge wissenschaftliche Mitarbeiter/-innen: Kompetenzprofil und Lehrveranstaltungsevaluation als Instrumente hochschulischer Personalentwicklung.* Dissertation, Friedrich-Schiller-Universität Jena.

Schmidt, B. & Loßnitzer, T. (2010). Lehrveranstaltungsevaluation: State of Art, ein Definitionsvorschlag und Entwicklungslinien. *Zeitschrift für Evaluation*, *9* , 49 – 72.

Sengewald, E. & Vetterlein, A. (2013). Gute Lehre kann man messen! Mehrebenen-Faktorenanalyse in der Lehrevaluation [Abstract]. In H. Weber (Hrsg.), *12. Arbeitstagung der Fachgruppe Differentielle Psychologie, Persönlichkeitspsychologie und Psychologische Diagnostik* (S. 115). Universität Greifswald.

Staufenbiel, T. (2000). Fragebogen zur Evaluation universitärer Lehrveranstaltungen durch Studierende und Lehrende. *Diagnostica*, *46*, 169 – 181.

Steyer, R., Mayer, A. & Fiege, C. (2014). Causal inference on total, direct, and indirect effects. In A. C. Michalos (Eds.), *Encyclopedia of Quality of Life and Well-Being Research* (pp. 606 – 630). Dordrecht: Springer.

Steyer, R., Mayer, A., Geiser, C. & Cole, D. (2015). A theory of states and traits–revised. *Annual Review of Clinical Psychology, 11*, 71 – 98. doi: 10.1146/annurev-clinpsy-032813-153719

Steyer, R., Schmidt, B. & Loßnitzer, T. (2004). *Situationsbericht 2004: zum Stand der Lehrveranstaltungsevaluation.* Friedrich-Schiller-Universität Jena.

Tufte, E. R. (2011). *The visual display of quantitative information.* Cheshire, CT: Graphics Press.

Wolbring, T. (2013). *Fallstricke der Lehrevaluation. Möglichkeiten und Grenzen der Messbarkeit von Lehrqualität.* Frankfurt am Main: Campus.

Dipl.-Psych. Anja Vetterlein
Dipl.-Psych. Erik Sengewald

Friedrich-Schiller-Universität Jena
Institut für Psychologie
Lehrstuhl für Methodenlehre und Evaluationsforschung
Am Steiger 3 / Haus 1
07743 Jena
E-Mail: anja.vetterlein@uni-jena.de

7 Abschlussdiskussion

Die Qualität der Lehre an Hochschulen ist ein Thema mit langer Tradition und wird bereits seit 1927 empirisch untersucht (vgl. Remmers & Brandenburg, 1927). Möchte man etwas über die Qualität einer Lehrveranstaltung erfahren, so ist ein Messinstrument erforderlich, das zuverlässig die relevanten Aspekte guter Lehre erfasst. Diese Aspekte festzulegen und sie zu messen, stellt in der Hochschulforschung bis heute eine wesentliche Herausforderung dar.

Die vorliegende Arbeit beschäftigt sich vorrangig mit der Methode zur Messung der Lehrveranstaltungsqualität. Hierfür werden in erster Linie Fragebögen eingesetzt, die ein vorher festgelegtes Spektrum an Facetten messen soll. Die Werte dieser Facetten werden dann im Sinne der Qualität der Lehrveranstaltung interpretiert. Die Fragebögen stehen jedoch häufig in der Kritik. Hauptsächlich wird dabei die psychometrische Qualität von LVE-Fragebögen in Frage gestellt, die überwiegend durch eine schlechte Modellpassung in konfirmatorischen Faktorenanalysen gekennzeichnet sind (vgl. Marsh et al., 2009; Ory, 2001; Spooren et al., 2013; Spooren, Mortelmans & Christiaens, 2014). Zum Teil wird den Fragebögen zur LVE ihr Nutzen als Messinstrument von Lehrqualität ganz abgesprochen und sie werden als Fragebögen zur Kundenzufriedenheit betrachtet (Beecham, 2009). Autoren, wie zum Beispiel Marsh et al. (2009) und Rindermann (2009) zeigen, dass LVE-Fragebögen häufig mit inadäquaten Verfahren bezüglich ihrer psychometrischen Qualität untersucht werden und die Standards konfirmatorischer Verfahren oft nicht erfüllen. Richardson (2005) betonen, dass die meisten Fragebögen zudem gar nicht erst auf ihre psychometrischen Eigenschaften hin untersucht werden. Ohne ein psychometrisch akzeptables Messmodell eines LVE-Fragebogens sind alle weiterführenden Diskussionen zur Validität der Evaluationsergebnisse hinfällig. Aber auf valide Ergebnisse, die im Sinne der Qualität einer Veranstaltung interpretiert werden können, ist der Dozent angewiesen. Beruhen die mittels Fragebogen gewonnenen Evaluationsergebnisse eher auf anderen Variablen anstatt auf der Qualität der Lehrveranstaltung, würde eine Veränderung der Lehre nicht zu einer messbaren Veränderung der Lehrveranstaltungsqualtät führen. Aufgrund der kritisierten psychometrischen Qualität der LVE-Fragebögen wird ihnen damit auch ihre Nützlichkeit abgesprochen, zur Verbesserung der Lehre beizutragen (vgl. Dresel & Rindermann, 2011; Marsh, 2007a). Neben der

Feedbackfunktion der LVE ist die Verbesserung der Lehre ein wesentliches Ziel der LVE. Auf Basis der Evaluationsergebnisse soll der Dozent die richtigen Schlüsse ableiten, Veränderungsbedarf erkennen und Maßnahmen einleiten. Anhand weiterer Evaluationen soll er dann überprüfen können, ob die Maßnahmen ihre intendierte Wirkung zeigen. Die Verbesserung der Lehre kann jedoch erst untersucht werden, wenn man die Qualität einer Lehrveranstaltung bzw. damit assoziierte Facetten messen kann. Dem Messmodell eines Fragebogens zur LVE wird daher besondere Aufmerksamkeit gewidmet.

Die erste Studie der vorliegenden Arbeit widmete sich zunächst der Fragestellung, ob für LVE-Fragebögen die richtigen Methoden zur Prüfung des Messmodells eingesetzt werden oder, ob neuere Verfahren wie die ML-CFA besser geeignet sind, um das theoretische Messmodell zu prüfen (vgl. Manuskript 1 in Kapitel 4). In der zweiten Studie wurde eine weitere mögliche Ursache für die zitierte schlechte Modellpassung von LVE-Fragebögen untersucht. Betrachtet wurden die strukturellen Eigenschaften der LVE-Gesamtstichproben, die zur Untersuchung der Passung der Messmodelle herangezogen werden. Einerseits betrifft dies Studenten, die mehr als eine Veranstaltung evaluieren (Mehrfachevaluation auf Studentenebene) und andererseits Dozenten, die mehr als eine Veranstaltung evaluieren lassen (Mehrfachevaluation auf Dozentenebene). Der Einfluss der Mehrfachevaluation auf Studenten- und Dozentenebene auf die CFA-Ergebnisse wurde bisher für den Kontext der LVE nicht untersucht.

Die Ergebnisse aus Studie 1 und Studie 2 hinterfragen die gängigen Praktiken zur Überprüfung des Messmodells von LVE-Fragebögen und geben wichtige Hinweise zur Verbesserung des aktuellen Standards. Am Beispiel des Fragebogens PELVE wird somit veranschaulicht, wie die Entscheidung über das Messmodell getroffen werden kann. Damit ist es möglich, den PELVE für weiterführende Fragestellungen zur Nützlichkeit der LVE für die Verbesserung der Lehre einzusetzen. Studie 3 zeigte eine Untersuchung der Nützlichkeit, wobei die Berichtform der LVE-Ergebnisse variiert und in einem experimentellen Design den Dozenten einer der beiden Berichte zugesandt wurde. Als Kriterium zur Untersuchung des Effekts der unterschiedlichen grafischen Darstellung wurden die LVE-Ergebnisse des Folgesemesters der Dozenten betrachtet. Als Maß für die Lehrveranstaltungsqualität bezüglich verschiedener Aspekte der Lehre wurden die Faktorwerte der latenten Variablen auf Veranstaltungsebene verwendet, die durch das Multilevel-Messmodel geschätzt wurden.

Im Folgenden werden die Ergebnisse der Studien kurz zusammengefasst, Einschränkungen diskutiert und der Beitrag für die Forschung skizziert. Das Kapitel schließt mit einer Diskussion zur Verwendbarkeit von LVE-Ergebnissen auf der Grundlage der empirischen Ergebnisse.

7.1 Messmodelle in der LVE

Der Vergleich verschiedener CFA-Verfahren zur Überprüfung des Messmodells eines Fragebogens zur LVE erfolgte am Beispiel des PELVE (Loßnitzer et al., 2007). Der Fragebogen PELVE wurde unter Beteiligung von Dozenten, Studenten und den Evaluationsverantwortlichen im Universitätsprojekt Lehrevaluation an der FSU Jena entwickelt und deckt einen breiten Bereich relevanter Facetten der Lehrveranstaltungsqualität ab (Loßnitzer et al., 2007; Schmidt & Loßnitzer, 2010). Durch die Kombination aus Prozess- und Ergebnissvariablen folgt der Fragebogen einer integrierten Theorie zur Lehrveranstaltungsevaluation, die die Qualität der Lehre als Zusammenspiel vielfältiger Faktoren betrachtet. Aus dieser Theorie des Fragebogens heraus lässt sich ein Messmodell formulieren, indem die Items des Fragebogens die jeweils postulierte Dimension konstruieren. Damit ist die Frage nach dem Messmodell des Fragebogens allerdings nicht beantwortet, weil die Passung des theoretisch postulierten Messmodells empirisch geprüft werden muss. Hierfür werden konfirmatorische Verfahren (CFA-Verfahren) eingesetzt, die in Studie 1 und Studie 2 Verwendung fanden. Die CFA-Verfahren unterscheiden sich insofern, als dass sie unterschiedliche Modelle testen weil verschiedene modellimplizierte Varianz-Kovarianz-Matrizen auf ihre Passung zur empirischen Varianz-Kovarianz-Matrix überprüft werden, obwohl die inhaltliche Zuordnung der Items zu latenten Variablen bei allen Verfahren gleich ist. In Studie 1 wurde zunächst die CFA auf Studentenebene betrachtet. Hier gehen die Antworten der Studierenden direkt in die Analyse des Messmodells ein. Die latenten Variablen in diesem Modell sind auf Studentenebene definiert und lassen keine unmittelbaren Aussagen über die Veranstaltung zu. Eine Theorie zur Erfassung der Veranstaltungsqualität kann mit diesem Verfahren nicht geprüft werden. Weiterhin wurde in Studie 1 die CFA auf Veranstaltungsebene untersucht. Hier wurden die in der Literatur berichteten schlechten Ergebnisse bezüglich der Modellpassung auch für den PELVE beobachtet. Als adäquates Modell stellt sich die Multilevel-CFA dar. Mit Hilfe der ML-CFA ist es möglich, die Gruppierung der Studenten in Lehrveranstaltungen bei der Faktorenanalyse zu berücksichtigen und gleichzeitig Varianzen und Kovarianzen latenter Variablen auf Within- und Between-Ebene zu identifizieren (vgl. B. O. Muthén, 1991). Toland und de Ayala (2005) betonen, dass bis zum Zeitpunkt ihrer Publikation wenige Forschungsarbeiten mit Multilevel-Analysen im Bereich der LVE gab (z. B. Marsh & Hattie, 2002; Ting, 2000) und keine Publikation vorliegt, die eine ML-CFA durchführt, um die Dimensionalität eines LVE-Fragebogens zu überprüfen. Nach den vorliegenden Ergebnissen der Studie 1 eignet sich die ML-CFA bereits auf Theorieebene zur Überprüfung des zugrunde liegenden Modells. Unter Verwendung der Studentenurteile können latente

Variablen auf Veranstaltungsebene konstruiert werden, die schließlich zur weiteren Analyse (vgl. Studie 3) verwendet werden können. Die ML-CFA zeigt in Studie 1 die besten Kennwerte für die Modellpassung. An dieser Stelle kann geschlussfolgert werden, dass die Kritik vieler Autoren zur schlechten Modellpassung der LVE-Fragebögen auf der Verwendung inadäquater Methoden zur Überprüfung der Modellpassung beruhen kann. Es empfiehlt sich in Anlehnung an die Ergebnisse aus Studie 1 die ML-CFA zur Prüfung des Messmodells eines LVE-Fragebogens in Betracht zu ziehen. Zuvor kann bereits auf der Theorieebene entschieden werden, ob eine ML-CFA angebracht ist und welche latenten Variablen auf Veranstaltungsebene betrachtet werden sollen. In der vorliegenden Studie 1 wurde anhand der Intraklassenkorrelation gezeigt, dass Varianz der abhängigen Variablen (die Items) durch die Veranstaltungszugehörigkeit aufgeklärt werden kann. Das unterstützt die Annahme, dass die Studentenurteile von Eigenschaften der Lehrveranstaltung abhängen. Studenten innerhalb einer Veranstaltung teilen ein Setting, das ihre Antworten im Fragebogen beeinflusst. Im Falle der LVE handelt es sich dabei um intendierte Effekte der Lehrveranstaltung auf die Studierenden. In der Evaluation sollen Studenten verschiedene Aspekte der Lehrveranstaltung einschätzen. Im Gegensatz zur Selbsteinschätzung ist es intendiert, dass die Fremdeinschätzung durch das zu bewertende Objekt beeinflusst wird. Sofern die Theorie des LVE-Fragebogens darauf ausgelegt ist, Aspekte oder Facetten der Lehrveranstaltung zu erheben, ist die Lehrveranstaltung als Analyseeinheit zu verwenden. Fragebögen zur LVE werden zum Teil auch als Selbstberichtverfahren eingesetzt. Hierbei geben die Studenten zum Beispiel Auskunft über ihren individuellen Kompetenzerwerb (vgl. Braun, 2008). Ob auch bei dieser Art der LVE-Instrumente eine ML-CFA geeignet ist, um das Messmodell des Fragebogens zu prüfen, muss im Einzelfall entschieden werden. In Studie 1 wurde deutlich, dass es sich bei der ML-CFA um ein Verfahren handelt, welches die Theorie der eingesetzten Fragebögen besser abbilden kann, als Verfahren, die die hierarchische Struktur der LVE-Daten nicht berücksichtigen. Neben der ML-CFA als modellbasiertes Verfahren, gibt es eine weitere Möglichkeit die hierarchische Struktur der LVE-Daten zu berücksichtigen. Die designbasierten Verfahren (vgl. Skinner et al., 1989; Wu & Kwok, 2012) eignen sich unter bestimmten Annahmen ebenso gut wie die ML-CFA zur Berücksichtigung des Mehrebenen-Designs bei der Analyse der Dimensionalität eines LVE-Fragebogens (Wu & Kwok, 2012; B. O. Muthén & Satorra, 1995). Dieses Verfahren wurde im Vorfeld der Studien der vorliegenden Arbeit ebenfalls berücksichtigt. Es zeigt eine ähnlich gute Modellpassung, fand jedoch aufgrund seiner Schwächen nicht Eingang in die Studien dieser Arbeit. So ist es erforderlich, dass bei der designbasierten Methode das Between-Modell mit dem Within-Modell strukturell identisch ist. Die Anzahl latenter Variablen ist zwar bei dem PELVE auf Within- und Between-Ebene gleich, jedoch

ermöglicht hier die ML-CFA eine flexible Anpassung des Modells und ist im Sinne der Modelloptimierung eher zu empfehlen. Bei der ML-CFA werden zudem die Ladungsparameter, Varianzen und Kovarianzen getrennt für beide Ebenen geschätzt und es existiert ein explizit formuliertes Within- und Between-Modell. Bei dem designbasierten Verfahren werden Korrekturformeln verwendet, um der hierarchischen Struktur Rechnung zu tragen. Es werden Standardfehler und Modellfit um den Einfluss der Gruppierung von Studenten in Veranstaltungen und demnach um den Einfluss der Varianz zwischen Veranstaltungen korrigiert (B. O. Muthén, 1994; L. K. Muthén & Muthén, 1998–2010). Die ML-CFA ermöglicht durch die explizite Trennung von Within- und Between-Modell eine größere Flexibilität und die Konstruktion latenter Variablen auf Veranstaltungsebene. Das Random-Intercept-Modell ermöglicht es, die latenten Variablen auf Veranstaltungsebene veranstaltungsspezifische Ursache für das Zustandekommen der Evaluationsergebnisse zu betrachten. Damit können die latenten Variablen auf Veranstaltungsebene zur Messung der Lehrveranstaltungsqualität herangezogen werden.

7.2 Mehrfachevaluation in der LVE

Ausgangspunkt und Motivation der Studie 1 war die vielmals als schlecht postulierte Modellpassung von Fragebögen zur LVE. Die Frage, ob eine schlechte Modellpassung durch die Wahl des CFA-Verfahrens zustande kommen kann, wurde in Studie 1 durch einen Vergleich verschiedener CFA-Verfahren untersucht. Im Ergebnis wurde deutlich, dass Verfahren, die die hierarchische Struktur der Stichproben berücksichtigen, eine bessere Modellpassung zeigen. Eine weitere mögliche Ursache für eine schlechte Modellpassung kann in der für die CFA verwendeten Stichprobe liegen. Studenten evaluieren während des Studiums mehrere Veranstaltungen desselben oder unterschiedlicher Dozenten. Ebenso lassen Dozenten mehrere Veranstaltungen evaluieren. Es ist zunächst plausibel anzunehmen, dass nicht nur die Lehrveranstaltung einen Einfluss auf das Urteil der Studenten hat, sondern die Studenten auch selbst unterschiedliche Antworttendenzen oder einen unterschiedlichen Maßstab der Bewertung haben. Studie 2 widmet sich aus diesem Grund der Frage nach dem Einfluss der Mehrfachevaluation auf die Überprüfung des Messmodells des PELVE-Fragebogens. Es wurden die drei CFA-Verfahren aus Studie 1 vergleichend gegenübergestellt und der Einfluss der Mehrfachevaluation auf die jeweiligen Ergebnisse der Modellgüte untersucht. Hierfür wurden verschiedene Stichprobentypen generiert, die eine Mehrfachevaluation auf Studenten- und/oder Dozentenebene beinhalteten. Die Feststellung der Mehrfachevaluation auf Dozenten- bzw. Studentenebene war in der vorliegenden Gesamtstichprobe möglich und ging der Generierung der unterschiedlichen Stichprobentypen voraus. Auf Dozentenebene diente der Name des Dozenten zur Identifikation der zugehörigen Veranstaltungen und damit verbunden der Mehrfachevaluation auf Dozentenebene. Die Veranstaltung wird durch den Dozenten selbst angemeldet und ist dadurch eindeutig zugeordnet. Allerdings können sich die Namen der Dozenten über die Zeit ändern. Zum Teil haben Dozenten ihr Zugangspasswort vergessen und daraufhin ein neues Benutzerkonto unter einem leicht geänderten Namen (z. B. mit gekürztem Vornamen) angelegt, in der Zwischenzeit promoviert und ihren Titel dem Namen hinzugefügt oder geheiratet und den Namen geändert. Bei einer automatischen Feststellung der Mehrfachevaluation über den Namen des Veranstaltungsleiters wäre ein vollständiger Ausschluss der Mehrfachevaluationen auf Dozentenebene nicht sicher gewesen, sodass die Namen zunächst manuell gesichtet und gegebenenfalls angeglichen wurden. Damit wird sichergestellt, dass jeder Dozent mit einem eindeutigen Namen in der Gesamtstichprobe vorhanden ist. Lediglich im Falle einer vollständigen Namensänderung durch zum Beispiel die Änderung des Familienstandes eines Dozenten kann es vorkommen, dass dieselbe Person mit verschiedenen Namen als Veranstal-

tungsleitung in der Stichprobe enthalten ist. Diese Fälle konnten nur in bekannten Einzelfällen identifiziert werden. Weitere Fälle sind hypothetisch in der Stichprobe enthalten, sollten sich jedoch auf ein Minimum beschränken. Auf Studentenebene erfolgte die Bereinigung der Gesamtstichprobe um die Mehrfachevaluation unter Verwendung des Personencodes (vgl. Studie 2 in Abschnitt 5). Studierende, die keinen Personencode angaben, wurden ebenfalls ausgeschlossen. Damit wurde für die Bereinigung der Stichprobe um die Mehrfachevaluation auf Studentenebene sichergestellt, dass auch unter Studenten ohne Personencode keine Mehrfachevaluation enthalten war.

Über 48 % der Studenten waren mit mehr als einer Evaluation in der Gesamtstichprobe, bestehend aus allen LVE seit 2005, enthalten. Diese Zahl kann als untere Schätzung der Mehrfachevaluation betrachtet werden. Der Anteil könnte höher sein, denn auch unter den Studenten, ohne die Angabe ihres Personencodes, kann Mehrfachevaluation auftreten und deren Anteil kann ggf. höher sein, wenn vor allem mehrfach evaluierende Studenten nicht möchten, dass ihre Daten über einen Personencode zusammengeführt werden können. Ist die Mehrfachevaluation in der Stichprobe, die der CFA zugrunde liegt, enthalten, hat dies im Vergleich zu einer Stichprobe ohne Mehrfachevaluation negative Folgen für die Kennwerte der Modellpassung. Dabei liegen Unterschiede zwischen den verschiedenen CFA-Verfahren hinsichtlich der Effekte vor. Während bei der CFA auf Veranstaltungsebene die Mehrfachevaluation auf Studentenebene maßgeblich zur Verschlechterung der Modellpassung beiträgt, hat diese Art der Mehrfachevaluation bei der ML-CFA keinen Einfluss auf die Modellpassung. Bei der ML-CFA schlägt sich hingegen eine mehrfache Evaluation auf Dozentenebene negativ auf die Modellpassung nieder. Studie 2 hat damit eine weitere Quelle (neben dem geeigneten CFA-Verfahren) für die oft zitierte schlechte Modellpassung von LVE-Fragebögen identifiziert. Besonders die oft verwendete CFA auf Veranstaltungsebene beansprucht für sich nicht sensitiv gegenüber der Mehrfachevaluation zu sein (vgl. Rindermann, 2009), zeigt in Studie 2 jedoch die geringste Robustheit gegenüber der Mehrfachevaluation auf Studentenebene. Autoren wie zum Beispiel Clayson (2009) verdeutlichen, dass in der Literatur die Analysen auf Veranstaltungsebene mit vorangehender Aggregation der Studentenurteile zu Veranstaltungsmittelwerten, die dominierende Analysemethode ist. Studie 2 zeigt, dass die Mehrfachevaluation negative Effekte auf die Modellpassung hat und durch die Aggregation die individuellen Einflüsse der Studenten nicht ausgemittelt werden. Sofern eine Identifikation der Mehrfachevaluationen möglich ist, lässt sich mit dem Vorgehen aus Studie 2 der Einfluss der Mehrfachevaluation untersuchen. Mit dem Ausschluss von Fällen geht allerdeings einher, dass kleine Veranstaltungen noch kleiner werden und der itemspezifische Veranstaltungsmittelwert nur noch mit einem größeren Standardfehler geschätzt werden kann. Für

die Analyse des Messmodells eines Fragebogens ist das zwar weniger problematisch, es ist jedoch empfehlenswert eine Methode zu identifizieren, die robust gegenüber der Mehrfachevaluation auf Studentenebene ist bzw. diese in das Modell einbezieht. In diesem Fall können alle vorhandenen Daten verwendet und in eine weiterführende Untersuchung einbezogen werden. Studie 2 zeigt, dass unter Ausschluss der Mehrfachevaluation starke Schwankungen der geschätzten Faktorwerte zu erwarten sind. Für weiterführende Untersuchungen zur Lehrveranstaltungsqualität, unter Verwendung der Faktorwerte, ist eine exakte Punktschätzung und eine stabile Schätzung derselben notwendig. Es bedarf daher eines Verfahrens, das robust gegenüber der Mehrfachevaluation eine gute Modellpassung zeigt und keine zusätzlichen Stichprobenfehler einführt, wie es bei dem Ausschluss der Mehrfachevaluation der Fall ist. Die ML-CFA zeigt in Studie 2 dahingehend gute Ergebnisse. Sie zeigt ein relativ robustes Verhalten bezüglich der Modellpassung, trotz der Mehrfachevaluation auf Studentenebene. Eine mögliche Ursache hierfür ist die Trennung von Within- und Between-Messmodell. Das Within-Messmodell auf Studentenebene beschreibt das Messmodell innerhalb von Veranstaltungen. In diesen kommt per Design keine Mehrfachevaluation vor. Mehrfachevaluation auf Studentenebene besteht nur zwischen verschiedenen Veranstaltungen. Damit ist ein Ausschluss der Mehrfachevaluation auf Studentenebene nicht notwendig, wenn die ML-CFA verwendet wird und die vollständigen Veranstaltungen in das Messmodell einfließen. Die ML-CFA reagiert lediglich empfindlich hinsichtlich der Kennwerte zur Modellpassung, wenn Mehrfachevaluation auf Dozentenebene enthalten ist. Gehen mehrere Veranstaltungen eines Dozenten in die Stichprobe ein, verschlechtern sich die Kennwerte der Modellpassung. Möglicherweise ist eine dritte Ebene im Modell nötig, die die Gruppierung von Veranstaltungen zu Dozenten abbildet. Eine derartige Dozentenebene ist dann relevant, wenn latente Variablen auf dieser Ebene konstruierbar sind. Im Fragebogen PELVE ist dies für die Facette *Dozentenverhalten* denkbar, weil diese Facette möglicherweise über verschiedene Veranstaltungen desselben Dozenten Ähnlichkeiten aufweist, die auf den Dozenten zurückführbar sind. Für andere Facetten, wie zum Beispiel *Rahmenbedingungen*, scheint dies weniger relevant zu sein.

Studie 2 zeigt damit, dass in Abhängigkeit des gewählten CFA-Verfahrens und der analysierten Stichprobe mit bzw. ohne Mehrfachevaluation, unterschiedliche Modellpassungen resultieren. Auf Basis der CFA-Ergebnisse wird meist über die Güte des Fragebogens geurteilt, sodass in Abhängigkeit des verwendeten CFA-Verfahrens unterschiedliche Urteile bezüglich der Güte des Fragebogens getroffen werden. So würde man unter Verwendung der CFA auf Veranstaltungsebene die Modellpassung und damit den PELVE-Fragebogen als schlecht und im Falle der ML-CFA als gut bewerten. Insbesondere die typische Konstellation aus der Verwendung der CFA auf Veranstaltungsebene und einer

Stichprobe mit Mehrfachevaluation auf Studentenebene würde den Eintruck hinterlassen, es handle sich um einen schlechten Fragebogen. Es erscheint plausibel, dass auch bei der CFA für andere Fragebögen zur LVE die Mehrfachevaluation eine mögliche Ursache für eine schlechte Modellpassung ist. Sofern in den analysierten Stichproben mehrere Veranstaltungen zu einer Gesamtstichprobe ohne Berücksichtigung der dadurch entstehenden Mehrfachevaluation zusammengeführt werden, entstehen stochastische Abhängigkeiten zwischen Beobachtungen aus verschiedenen Veranstaltungen, die von den konventionellen CFA-Verfahren nicht berücksichtigt werden.

7.3 Rezeption von LVE-Ergebnissen

In Studie 1 und Studie 2 hat sich die ML-CFA als geeignetes Verfahren herauskristallisiert, um das Messmodell des Fragebogens PELVE adäquat abzubilden. Studie 3 verwendet das Messmodell und untersuchte, inwiefern die Variantion der Ergebnisdarstellung im Ergebnisbericht einen Einfluss auf die Lehrveranstaltungsqualität hat. Hierfür wurden die latenten Variablen auf Veranstaltungsebene als entsprechende Facette der Lehrveranstaltungsqualität betrachtet und deren Abhängigkeit von der Ergebnisdarstellung als Gruppenvariable in einem randomisierten Experiment untersucht. In Studie 3 wurde eine Manipulation der Ergebnisrezeption angestrebt. Die Rezeption der Ergebnisse durch den Dozenten lässt sich nicht direkt manipulieren. Vielmehr müssen indirekte Wege gefunden werden, die Rezeption der Ergebnisse zu verbessern, sodass ein Veränderungsbedarf möglichst leicht erkannt wird. Ein Weg, Einfluss auf die Rezeption zu nehmen, ist die Änderung der Ergebnisdarstellung. Während die meisten LVE-Instrumente oder -Designs keine Möglichkeit zur Diskussion der Ergebnisse mit den Studenten ermöglichen (Johnson, 2000), gestattet der LVE-Prozess an der FSU Jena (vgl. Kapitel 2.2.3) ein zeitnahes Feedback der Ergebnisse in einer Bericht-, Aushang- und Präsentationsform. Dies verdeutlicht, dass der Dialog über die LVE-Ergebnisse zwischen Dozenten und Studenten ein wichtiger Bestandteil der LVE an der FSU Jena ist. Folglich wird auf die Aufbereitung der Ergebnisse und speziell auf grafische Darstellungen besonders viel Wert gelegt und eine Manipulation derselben stets durch empirische Forschung begleitet. Ziel der Studie 3 war es, die Überarbeitung der Ergebnisgrafiken in Bezug auf mögliche Effekte auf die Lehrveranstaltungsqualität empirisch zu begleiten. Die Datengrafiken sollten überarbeitet und im Vergleich zur bisherigen Darstellung kompakter werden, um diese direkt in Kombination mit dem Itemtext und den deskriptiven Kennwerten in einer gemeinsamen Tabelle darzubieten. Durch die Integration der Antwortverteilung in die grafische Darstellung der deskriptiven Kennwerte auf Itemebene, wurde die Grafik komplexer als es bisher

der Fall war (vgl. Studie 3 in Abschnitt 6). Gleichzeitig wurde die Länge des Berichts durch diese Maßnahme halbiert. Solch massive Änderungen an der Aufbereitung der Evaluationsergebnisse können sich auf den Rezeptionsprozess auswirken. Wenn die komplexere Grafik von den Dozenten missverstanden wird, sollte von der geplanten Änderung abgesehen werden. Aus diesem Grund wurde die Überarbeitung des Ergebnisberichts durch verschiedene Vorstudien (vgl. Abschnitt 2.2.3) und das in Studie 3 geschilderte Experiment begleitet. Die Evaluation erfolgte in Studie 3 anhand eines distalen Kriteriums (den LVE-Ergebnissen der Folgeveranstaltung). Für diese Fragestellung postuliert Marsh (2007a) keine Veränderung über die Zeit, weil nur der Evaluationsbericht als Feedback verwendet wird. Vielmehr müssten die Dozenten konsolidiert werden, um die Ergebnisse mit ihnen zu besprechen. Dresel und Rindermann (2011) finden für ein derartiges Vorgehen mittlere bis große Effekte ($d = .68$). In einer der Vorstudien (vgl. Abschnitt 2.2.3) wurden Interviews mit den Dozenten durchgeführt, wobei ihnen die Ergebnisse genau erläutert wurden. Diese Art der Ergebnisrückmeldung ist nur schwer in das Tagesgeschäft der LVE zu integrieren und lässt sich nur mit hohem Personalaufwand realisieren. Mit bis zu 700 evaluierten Veranstaltungen an der FSU Jena je Semester müssen ökonomische Methoden, wie die Manipulation der Ergebnisberichte, eingesetzt werden, um das Feedback zu manipulieren. In Studie 3 wurden daher die grafischen Elemente derart verändert, dass eine kompakte und gleichzeitig informative Ergebnisgrafik resultierte. Der neue Ergebnisbericht wurde anschließend in das Tagesgeschäft der LVE eingeführt und für eine zufällige Auswahl an Dozenten versendet.

Die Ergebnisse der Studie 3 deuten darauf hin, dass die Manipulationen des Ergebnisberichts Konsequenzen für die Lehrveranstaltungsqualität haben können. Die kleinen Effekte in Studie 3 replizieren vergleichbare Ergebnisse von Lang und Kersting (2007) und Marsh (2007a). Eine mögliche Ursache für die kleinen Effekte kann die randomisierte Zuweisung zu den Ergebnisberichten sein, wobei die Notwendigkeit zur Verbesserung der Lehre unberücksichtigt blieb. Durch die Randomisierung können auf theoretischer Ebene differenziell wirkende Einflussgrößen ausgeschlossen werden. Damit kann angenommen werden, dass der Veränderungsbedarf bezüglich der Lehre in beiden Gruppen gleich ist. Ist der Veränderungsbedarf in beiden Gruppen niedrig, sind nur wenige Änderungen seitens der Dozenten zu erwarten. Die Notwendigkeit zur Verbesserung muss jedoch gegeben sein, damit eine positive Veränderung der LVE-Ergebnisse erwartet werden kann. Selbst wenn die Rezeption durch die kompakte Darstellungsform erleichtert wird, müssen die Ergebnisse keinen handlungsleitenden Aufforderungscharakter haben. Für weitere Studien empfiehlt es sich daher, vorab die Lehrenden bzw. die Veranstaltungen zu identifizieren, für die ein Veränderungsbedarf besteht. Ein Vergleich verschiedener Ergebnisdarstellungen für Gruppen mit Veränderungsbedarf wäre eine

geeignete Studie zur Untersuchung des Effektes von verschiedenen Ergebnisdarstellungen auf die Ergebnisrezeption und die Veränderung der Lehrqualität. Als Fazit aus der dritten Studie ist die Notwendigkeit der empirischen Begleitung von Veränderungen der Ergebnisberichte zu ziehen. Genaues Wissen über die Art und Weise, wie die Evaluationsberichte gelesen werden und welche Informationen als relevantes Feedback erachtet werden bzw. welche Schlüsse gezogen werden, ist weitesgehend unbekannt. Änderungen am Ergebnisbericht können nicht-intendierte Effekte haben, die es zu vermeiden gilt. Durch die empirische Untersuchung solcher Entwicklungen kann der Feedbackprozess unterstützt und die Nützlichkeit der LVE belegt werden. Die Studie 3 bietet eine gute Vorlage, wie man eine solche empirische Untersuchung gestaltet und wie die bisherigen Ergebnisse zur Analyse der LVE-Daten im Rahmen eines ML-CFA Verfahrens unter Berücksichtigung von Mehrfachevaluationen angewendet werden.

7.4 Hinweise zur Verwendung von LVE-Ergebnissen

Während sich die vorherigen Abschnitte mit der Diskussion und Einordnung der Ergebnisse aus den Studien beschäftigten, soll dieser Abschnitt als kurzes Statement zur Lehrveranstaltungsevaluation und den aktuellen Entwicklungen betrachtet werden.

Lehrveranstaltungsevaluationen werden an deutschen Hochschulen zunehmend für mehrere Zwecke eingesetzt. Die vorrangige Funktion der LVE als Feedbackinstrument für den Dozenten ist nach wie vor eine wichtige. Der Lehrende setzt die LVE ein, um ein standardisiertes Feedback von seinen Studenten zu erhalten und Stärken und Schwächen seiner Veranstaltung auszumachen. Evaluationsergebnisse gehören daher vor allem in die Hände des Dozenten und sind idealerweise so aufbereitet, dass sie den Dialog mit den Studenten unterstützen. Eine derart ausgerichtete LVE ist als formative Evaluation angelegt und soll zur Verbesserung der Lehre beitragen. Eine anderweitige Verwendung der Evaluationsergebnisse birgt hingegen die Gefahr, dass die Feedbackfunktion der LVE verloren geht. So steht die zu beobachtende Verwendung der LVE-Ergebnisse als Instrument summativer Evaluation der Feedbackfunktion im Weg. Die LVE sollen die Lehrqualität über die konkrete Veranstaltung hinaus abbilden, für personelle Entscheidungen herangezogen und für institutionelle Audits verwendet werden (vgl. Arthur, 2009; Burden, 2008; Edström, 2008; Emery, Kramer & Tian, 2003). Für die Verwendung als vergleichendes Instrument sind jedoch objektivere Verfahren von hoher psychometrischer Qualität erforderlich, die den Einfluss potentieller Drittvariablen berücksichtigen, um

einen fairen Vergleich zwischen unterschiedlichen Evaluationen zu ermöglichen. Die LVE-Ergebnisse müssen zweifelsfrei der Lehrveranstaltung als genuiner Wirkfaktor zugeschrieben werden, d.h. es darf keinen Zweifel an der Reliabilität und Validität des Verfahrens geben, wie es bei LVE-Fragebögen der Fall ist (vgl. Aleamoni, 1999; Marsh, 1984, 1987, 2007b; Marsh & Roche, 1997). Aleamoni (1999), Centra (2003) und Marsh und Roche (2000) zeigen zwar, dass andere Einflussfaktoren einen relativ geringen Einfluss auf die LVE-Ergebnisse haben, dennoch ist unbekannt, welcher Anteil der Unterschiede in LVE-Ergebnissen durch vom Dozenten nicht beeinflussbare Variablen erklärt werden kann. Trotz dieser Unsicherheit bezüglich der Vergleichbarkeit unterschiedlicher LVE-Ergebnisse, werden globale Maße zum Ranking der Lehrveranstaltungsqualität verlangt (vgl. Apodaca & Grad, 2005; Spooren et al., 2013). Für die Dozenten wird die LVE damit zum unangenehmen Test der eigenen Lehrkompetenz, sodass die Tendenz beobachtet werden kann, dass Dozenten die Verbesserung ihrer LVE-Scores anstreben, anstatt eine Verbesserung der Lehre (Simpson & Siguaw, 2000). Die Frage: „Wie kann ich meine Evaluationsergebnisse verbessern?“ wird immer häufiger gestellt und ist der Frage: „Wie kann ich meine Lehre verbessern?“ vorangestellt. Eine derartige Entwicklung kann negative Effekte auf den Feedbackcharakter haben, den die LVE bisher in Anspruch nimmt. Die Nützlichkeit der LVE für die Verwendung auf administrativer Ebene als Kontroll- oder Managementinstrument wurde in der vorliegenden Arbeit nicht direkt untersucht. Jedoch deutet Studie 2 darauf hin, dass die LVE-Ergebnisse keine zuverlässige Grundlage für diese Funktion darstellen. Die Faktorwerte und damit die Schätzung der Lehrveranstaltungsqualität im Hinblick auf die enthaltenden Facetten, ändern sich in Abhängigkeit des gewählten CFA-Verfahrens und der Mehrfachevaluation, sodass das Ranking der Veranstaltungen variabel ist. Die hohe Stichprobenabhängigkeit eines Rankings liegt auf der Hand. Studie 2 untersucht ganz bewusst Prozentrangdifferenzen, wohl wissend, dass damit noch keine Aussagen über die statistische Signifikanz der Unterschiede gegeben ist und die Standardfehler der Faktorwertschätzung zu berücksichtigen wären. Auf Steuerungsebene werden die Rankings jedoch ebenfalls hauptsächlich in absoluten Werten betrachtet und ohne Berücksichtigung eines Konfidenzintervalls interpretiert. Die Analyse der Faktorwerte in Studie 2 soll für diesen Fall sensibilisieren und ein Alarmsignal für die Verwendung der LVE-Ergebnisse auf Steuerungsebene darstellen.

Für die Feedbackfunktion der LVE ist ein Ranking hingegen nicht notwendig. Um dennoch eine Einordnung in Relation zu anderen Dozenten und Veranstaltungen zu erhalten, kann eine grobe Verankerung der eigenen LVE-Ergebnisse in Relation zu Vergleichswerten im Bericht abgedruckt werden. Die Vergleiche zum Durchschnittswert der gesamten Hochschule und zum Durchschnittswert des Fachbereichs liefern zwar ebenfalls keine kausal interpretierbaren Differenzen, sie sind in der

Ergebnisdiskussion dennoch nützlich. Die Vergleichswerte liefern eine grobe Orientierung und suggerieren keine feingliedrige Differenzierbarkeit, wie es bei einem Ranking aller Veranstaltungen der Fall ist. Eine weitere wertvolle Unterstützung der Feedbackfunktion ergibt sich aus dem Vergleich der LVE-Ergebnisse der Studenten zur Perspektive des Dozenten. Der Dozent erhält während der Evaluation ebenfalls einen Fragebogen (vgl. Anhang A.3) und beantwortet die Items aus seiner Sicht, sodass ein Ist-Soll-Vergleich zwischen Dozenten- und Studentenmeinung möglich ist. Im Ergebnisbericht können die unterschiedlichen Perspektiven durch den Dozentenwert und den Durchschnittswert der Studentenurteile verglichen werden. Diese Art Feedback wird auch von Spooren et al. (2013) vorgeschlagen und soll die Transparenz der unterschiedlichen Perspektiven erhöhen und Diskussionen über verschiedene Vorstellungen von guter Lehre anregen.

Weitere Forschung ist nötig um den Nutzen von LVE für den Dozenten und die Studenten zu untersuchen. Dabei kann die Verwendung von ML-CFA helfen, die Erhebung der Lehrveranstaltungsqualität genauer abzubilden, als konventionelle Verfahren. Weiterführende Fragestellungen, wie die Entwicklung der Lehrveranstaltungsqualität über die Zeit oder die Effekte von hochschuldidaktischen Maßnahmen zum Training der Lehrkompetenz können somit genauer untersucht werden. Dabei ist jedoch zu beachten, dass die LVE nur ein Maß von vielen möglichen ist und nur einen kleinen Ausschnitt der qualitätsrelevanten Merkmale einer Lehrveranstaltung erfassen kann.

Literaturverzeichnis

Abrami, P. C. (1985). Dimensions of effective college instruction. *The Review of Higher Education, 8* (3), 211–228.

Abrami, P. C., d' Appolonia, S. & Cohen, P. A. (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology, 82* (2), 219–231.

Abramson, T. (1979). *Handbook of vocational education evaluation.* Beverly Hills: Sage.

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education, 13* (2), 153–166.

Apodaca, P. & Grad, H. (2005). The dimensionality of student ratings of teaching: Integration of uni- and multidimensional models. *Studies in Higher Education, 30*, 723–748.

Arthur, L. (2009). From performativity to professionalism: Lecturer's responses to student feedback. *Teaching in Higher Education, 14*, 441–454.

Astleitner, H. (1991). Studentische Einschätzungen von universitärem Lernverhalten: Das Problem impliziter Theorien. *Psychologie in Erziehung und Unterricht, 38*, 116–122.

Astleitner, H. & Krumm, V. (1996). Dimensionen von Lehrverhalten: Faktorenstrukturen 1. und 2. Ordnung mit Kreuzvalidierung. *Empirische Pädagogik, 10* (1), 7–26.

Basler, H., Bolm, G., Dickescheid, T. & Herda, C. (1995). Marburger Fragebogen zur Akzeptanz der Lehre. *Diagnostica, 4* (1), 62–79.

Beecham, R. (2009). Teaching quality and student satisfaction: Nexus or simulacrum? *London Review of Education, 7*, 135–146.

Beisteiner, A. (1999). *Lehrevaluation an der Universität Wien* (Bericht). Universität Wien.

Born, S., Loßnitzer, T. & Schmidt, B. (2006). Lehrveranstaltungsevaluation an der Friedrich-Schiller-Universität Jena - Eine Analyse der Dimensionalität der eingesetzten Fragebögen. In B. Krause & P. Metzler (Hrsg.), *Empirische Evaluationsmethoden* (Bd. 10, S. 99–116). Berlin: ZeE Verlag.

Braun, E. (2008). *Das Berliner Evaluationsinstrument für selbsteingeschätzte studentische Kompetenzen: Ein Lehrveranstaltungs-Evaluationsinstrument zur Erfassung des subjektiven Kompetenzerwerbs in Folge eines Lehrveranstaltungsbesuches.* V&R unipress.

Buhl, T. (1999). *Entwicklung eines Fragebogens zur Evaluation von Lehrveranstaltungen. Pilotphase im Wintersemester 1998/99 an der Friedrich-Schiller-Universität Jena.* (Bericht). Friedrich-Schiller-Universität Jena.

Burden, P. (2008). Does the end of semester evaluation forms represent teacher's views of teaching in a tertiary education contest in japan? *Teaching and Teacher Education, 24*, 1463–1475.

Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education, 44* (5), 495–518.

Clayson, D. E. (2006). Grades and the student evaluation of instruction: A test of the reciprocity effect. *Academy of Management Learning & Education, 5* (1), 52–65.

Clayson, D. E. (2007). Conceptual and statistical problems of using between-class data in educational research. *Journal of Marketing Education, 29*, 34–38.

Clayson, D. E. (2009). Student evaluations of teaching. Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education, 31*, 16–30.

Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. *Research in Higher Education, 13* (4), 321–341.

Diehl, J. M. & Kohr, H. (1977). Entwicklung eines Fragebogens zur Beurteilung von Hochschulveranstaltungen im Fach Psychologie. *Psychologie in Erziehung und Unterricht*, *24*, 61–75.

Dresel, M. & Rindermann, H. (2011). Counseling university instructors based on student evaluations of their teaching effectiveness: A multilevel test of its effectiveness under consideration of bias and unfairness variables. *Research in Higher Education, 52*, 1–21.

Edström, K. (2008). Doing course evaluation as if earning matters most. *Higher Education Research & Development*, *27*, 95–106.

Elbing, E., Gräsel, C. & Perleth, C. (1997). Workshop zur studentischen Lehrevaluation (Vorlesungsfragebogen MILVA). München: Kolloquium Pädagogische Psychologie.

Emery, C. R., Kramer, T. R. & Tian, R. (2003). Return to academic standards: A critique of students' evaluations of teaching effectiveness. *Quality Assurance in Education, 11*, 37–47.

Esser, H. (1994). Lehrbericht der Fakultät für Sozialwissenschaften der Universität Mannheim. Ergebnisse der Studenten- und Lehrerhebung im Wintersemester 1993/94. Mannheim: Fakultätsbericht.

Feldman, K. A. (1976). The superior college teacher from the students' view. *Research in Higher Education, 5*, 243–288.

Friedrich-Schiller-Universität Jena. (2012). Evaluationsstandards und Instrumente der Qualitätsent-

wicklung im Bereich Studium und Lehre (Evaluationsordnung). *Verkündungsblatt* (8/2012), 252–255.

Gold, A. & Mayring, P. (1997). *Dimensionen studentischer Seminarbewertung.* Poster auf der 6. Tagung Pädagogische Psychologie. Frankfurt.

Gollwitzer, M. & Schlotz, W. (2003). Das Trierer Inventar zur Lehrveranstaltungsevaluation (TRIL): Entwicklung und erste testtheoretische Erprobungen. In G. Krampen & H. Zayer (Hrsg.), *Psychologiedidaktik und Evaluation IV* (S. 114–128). Bonn: Deutscher Psychologen Verlag.

Hejj, A. (1999). *Wenn Wunschdozenten gewählt würden. Ein empirischer Beitrag zu rEvaluation von Hochschullehrern.* (Unveröffentlichtes Manuskript München: Institut für Psychologie)

Hofmann, J. M. (1990). Die Beurteilung pädagogisch-psychologischer Lehrveranstaltungen anhand des VB-PSYCH. *Psychologie in Erziehung und Unterricht*, *37*, 47–53.

Johnson, R. (2000). The authority of the student evaluation questionnaire. *Teaching in Higher Education*, *5*, 419–434.

Kleine, D. & Merkens, H. (1979). Überprüfung eines Fragebogens zur Beurteilung von Lehrveranstaltungen. *Psychologie in Erziehung und Unterricht*, *26*, 149–153.

Koch, E. (2004). *Gute Hochschullehre. Theoriebezogene Herleitung und empirische Erfassung relevanter Lehraspekte.* Hamburg: Verlag Dr. Kovač.

Kramis, J. (1990). Bedeutsamkeit, Effizienz, Lernklima. Grundlegende Gütekriterien für Unterricht und didaktische Prinzipien. *Beiträge zur Lehrerbildung*, *8*, 279–296.

Lang, J. W. B. & Kersting, M. (2007). Regular feedback from student ratings of instruction: Do college teachers improve their ratings in the long run? *Instructional Science*, *35*, 187–205.

Linn, R. L., Centra, J. A. & Tucker, L. (1975). Between, within, and total group factor analyses of student ratings of instruction. *Multivariate Behavioral Research*, *10*, 277–288.

Loßnitzer, T., Schmidt, B. & Born, S. (2007). Zentrale Lehrveranstaltungsevaluation an der Friedrich-Schiller-Universität Jena - Qualitätsmodell und Messinstrument. In M. Krämer, S. Preiser & K. Brusdeylins (Hrsg.), *Psychologiedidaktik und Evaluation VI.* (S. 327–335). Göttingen: V&R unipress.

Marsh, H. W. (1982a). SEEQ: A reliable, valid, and useful instrument for collecting students'evaluations of university teaching. *British Journal of Educational Psychology*, *52* (1), 77-92.

Marsh, H. W. (1982b). The use of path analysis to estimate teacher and course effects in student ratings of instructional effectiveness. *Applied Psychological Measurement*, *6* (1), 47–59.

Marsh, H. W. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. *Journal of Educational Psychology, 75* (1), 150–166.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76* (5), 707 – 754.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11*, 253–388.

Marsh, H. W. (2007a). Do university teachers become more effective with experience? A multilevel growth model of students' evaluations of teaching over 13 years. *Journal of Educational Psychology, 99* (4), 775–790.

Marsh, H. W. (2007b). Students' evaluations of university teaching: Dimensionality, reliablity, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (Hrsg.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (S. 319–383). Dordrecht: Springer.

Marsh, H. W. & Hattie, J. (2002). The relation between research productivity and teaching effectiveness. Complementary, antagonistic, or independent constructs? *Journal of Higher Education, 73* (5), 603–641.

Marsh, H. W., Muthén, B. O., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S. & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling, 16*, 439–476.

Marsh, H. W. & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 53*, 1187–1197.

Marsh, H. W. & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, or innocent bystanders? *Journal of Educational Psychology, 92* (1), 202–228.

Müller-Wolf, H.-M. (1977). *Lehrverhalten an der Hochschule.* München: Verlag Dokumentation.

Moosbrugger, H. & Schermeleh-Engel, K. (2006). Faktorenanalyse. In F. Petermann & M. Eid (Hrsg.), *Handbuch der Psychologischen Diagnostik* (S. 304–317). Hogrefe.

Multrus, F. (1995). *Zur Lehr- und Studienqualität. Dimensionen, Skalen und Befunde des Studierendensurveys.* Konstanz: Hefte zur Bildungs- und Hochschulforschung (12).

Murray, H. G. (1983). Low-inference classroom teaching behaviors and student ratings of college

teaching effectiveness. *Journal of Educational Psychology, 75* (1), 138–149.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28* (4), 338–354.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22* (3), 376–398. doi: 10.1177/0049124194022003006

Muthén, B. O. & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology, 25*, 267–316.

Muthén, L. K. & Muthén, B. O. (1998–2010). Examples: Multilevel modeling with complex survey data. In *Mplus user's guide* (Bd. 6). Los Angeles, CA: Muthén & Muthén.

Ory, J. C. (2001). Faculty thoughts and concerns about sutudent ratings. *New Directions for Teaching and Learning, 87*, 3–15.

Remmers, H. & Brandenburg, G. (1927). Experimental data on the purdue rating scale for instruction. *Educational Administration and Supervision, 13*, 519–527.

Richardson, J. T. E. (2005). Instruments for obtaining student feedback: A review of the literature. *Assessment and Evaluation in Higher Education, 30*, 387–415.

Rindermann, H. (2009). *Lehrevaluation: Einführung und Überblick zu Forschung und Praxis der Lehrveranstaltungsevaluation an Hochschulen mit einem Beitrag zur Evaluation computerbasierten Unterrichts* (2. Aufl.). Landau: Empirische Pädagogik e. V.

Ronning, R. R. & Walsh, U. R. (1977). Effects of student anonymity-nonanonymity on the factor structure of a teacher rating form. *Research in Higher Education, 6*, 363–371.

Schmidt, B. & Loßnitzer, T. (2010). Lehrveranstaltungsevaluation: State of Art, ein Definitionsvorschlag und Entwicklungslinien. *Zeitschrift für Evaluation, 9* (1), 49–72.

Scriven, N. (1972). Die Methodologie der Evaluation. In C. Wulf (Hrsg.), *Evaluation.* München: Piper.

Simpson, P. M. & Siguaw, J. A. (2000). Student evaluations of teaching: An exploratory study of the faculty response. *Journal of Marketing Education, 22*, 199–213.

Skinner, C. J., Holt, D. & Smith, T. M. F. (Hrsg.). (1989). *Analysis of complex surveys.* West Sussex, England: Wiley.

Spiel, C. & Gössler, P. (1998). *Sinn und Unsinn einer vergleichenden Evaluierung universitärer Lehre durch Studierende.* (Graz: unveröff. Manuskript)

Spooren, P. (2012). *The unbearable lightness of student evaluations of teaching in higher education* (Unveröffentlichte Dissertation). University of Antwerp, Belgium.

Spooren, P., Brockx, B. & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The

state of the art. *Review of Educational Research, 83* (4), 598–642.

Spooren, P., Mortelmans, D. & Christiaens, W. (2014). Assessing the validity and reliability of a quick scan for student's evaluation of teaching. Results from confirmatory factor analysis and G Theory. *Studies in Educational Evaluation.*

Staufenbiel, T. (2000). Fragebogen zur Evaluation von universitären Lehrveranstaltungen durch Studierende und Lehrende. *Diagnostica, 46* (4), 169–181.

Steyer, R., Mayer, A., Geiser, C. & Cole, D. A. (2015). A Theory of States and Traits – Revised. *Annual Review of Clinical Psychology, 11*, 1–28.

Suchman, E. A. (1967). *Evaluative research: Principle and practice in public service and social action programs.* Russell Sage Foundation.

Tetenbaum, T. (1977). The factor invariance of student ratings of instruction under three sets of directions. *Research in Higher Education, 6*, 11–23.

Thiel, F., Blüthmann, I. & Watermann, R. (2012). Konstruktion eines Fragebogens zur Erfassung der Lehrkompetenz (LeKo). In B. Berendt, H.-P. Voss & J. Wildt (Hrsg.), *Neues Handbuch Hochschullehre.* Stuttgart: Raabe Verlag.

Ting, K. F. (2000). A multilevel perspective on student ratings of instruction: Lessons from the chinese experience. *Research in Higher Education, 41* (5), 637–661.

Toland, M. D. & de Ayala, R. J. (2005). A multilevel factor analysis of students' evaluations of teaching. *Educational and Psychological Measurement, 65* (2), 272–296.

Vetterlein, A. & Sengewald, E. (2011). Mittelwerte verstehen! Rezeption von Ergebnisberichten im Hochschulsektor. In *Vortrag auf der 10. Tagung der Fachgruppe Methoden und Evaluation.* Bamberg.

Vetterlein, A. & Sengewald, E. (2015). Ergebnisdarstellung in der Lehrveranstaltungsevaluation. Effekte verschiedener Berichte auf die Qualität von Lehrveranstaltungen. *Diagnostica, 61*, 153–162.

Westermann, R., Spies, K., Heise, E. & Wollburg-Claar, S. (1998). Bewertung von Lehrveranstaltungen und Studienbedingungen durch Studierende: Theorieorientierte Entwicklung von Fragebögen. *Empirische Pädagogik, 12* (2), 133–166.

Winteler, A. & Schmolck, P. (1979). Entwicklung und Validierung eines Schätzverfahrens zur Beurteilung von Lehrveranstaltungen. *Schweizerische Zeitschrift für Psychologie, 38* (2), 139–156.

Winteler, A. & Schmolck, P. (1983). Überprüfung eines Schätzverfahrens zur Beurteilung von Lehrveranstaltungen. *Schweizerische Zeitschrift für Psychologie, 42* (1), 56–79.

Wittmann, W. (1985). *Evaluationsforschung. Aufgaben, Probleme & Anwendungen.* Berlin: Springer.

Wolbring, T. (2013). *Fallstricke der Lehrevaluation. Möglichkeiten und Grenzen der Messbarkeit von Lehrqualität.* Frankfurt am Main: Campus Verlag.

Wottawa, H. (1986). Evaluation. In B. Weidenmann, A. Krapp, M. Hofer, G. L. Haber & H. Mandl (Hrsg.), *Pädagogische Psychologie* (S. 703–733). Urban & Schwarzenberg.

Wottawa, H. & Thierau, H. (2003). *Lehrbuch Evaluation* (3. Aufl.). Bern: Hans Huber.

Wu, J. & Kwok, O. (2012). Using SEM to analyze complex survey data: A comparison between design-based single-level and model-based multilevel approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, *19* (1), 16–35.

A PELVE Fragebögen

Auf den folgenden Seiten dieses Anhangs sind die drei Fragebogenversionen des PELVE eingebunden. Dabei handelt es sich um den Fragebogen für Vorlesungen, Seminare und Übungen. Für jede Fragebogenversion liegt ein Fragebogen für Studenten und einer für Dozenten vor.

Abbildung A.1: Studentenfragebogen für Vorlesungen (Seite 1)

seit 1558

Friedrich-Schiller-Universität Jena

Universitätsprojekt Lehrevaluation

Lehrveranstaltungsevaluation

- Fragebogen für Vorträge und Vorlesungen (Teilnehmende) -

Mit diesem Fragebogen können Sie Ihren **persönlichen Eindruck** von dieser Lehrveranstaltung zurückmelden. Füllen Sie die zutreffende Antwortalternative bitte **mit einem dunklen Stift** (kein Bleistift) aus. Wenn eine Frage **nicht beantwortbar** ist oder Sie **keine Antwort geben möchten**, markieren Sie '**keine Angabe**' (k.A.). Möchten Sie eine **falsch markierte Antwort korrigieren,** streichen Sie diese bitte durch und markieren die von Ihnen gewünschte Antwortalternative. Die Fragebögen werden durch das Universitätsprojekt **Lehrevaluation (www.ule.uni-jena.de)** statistisch ausgewertet.

Veranstaltung **Dozent/-in** **Datum**

Bitte machen Sie für statistische Zwecke die folgenden Angaben.

Alter in Jahren ☐☐ Geschlecht ○ weiblich ○ männlich Fachsemester ☐☐

Bitte nennen Sie die Hauptgründe Ihres Veranstaltungsbesuches. (*Mehrfachnennungen möglich*)

○ inhaltliches Interesse ○ Pflichtveranstaltung ○ guter Ruf der Lehrkraft
○ keine Alternative verfügbar ○ zur Vorbereitung auf die Prüfung ○ andere Gründe

Ausschließlich für wissenschaftliche Fragestellungen bitten wir Sie, einen 5-stelligen Personencode anzugeben.

Der Personencode wird aus dem ersten Buchstaben Geburtsorts, dem zweiten Buchstaben Ihres Vornamens und dem dritten Buchstaben Ihres Nachnamens (ggf. Mädchenname) sowie den jeweils letzten Ziffern Ihres Geburtstags und Ihres Geburtsmonats gebildet. So lautet der Code für die aus Halle (1. Stelle) stammende Anja Krüger (2. und 3. Stelle), geb. am 26.03.1973 (4. und 5. Stelle des Codes): **HNÜ63**. ☐☐☐☐☐

An wievielen der bisherigen Termine dieser Veranstaltung haben Sie teilgenommen? . . . . . 0-20% ○ 21-40% ○ 41-60% ○ 61-80% ○ 81-100% ○

Wieviele **Stunden pro Woche** verbringen Sie durchschnittlich mit dem **Selbststudium** (bezogen auf alle Veranstaltungen **in diesem Semester**)? ☐☐,☐

Wieviele Stunden hiervon entfallen pro Woche auf **diese Veranstaltung**? ☐☐,☐

	stimme nicht zu			stimme zu		k.A.
Ich empfinde den von mir für diese Veranstaltung zu erbringenden Arbeitsaufwand als angemessen.	○	○	○	○	○	☐

5914529285

Bitte treffen Sie zunächst einige zusammenfassende Einschätzungen.	stimme nicht zu			stimme zu		k.A.
Die Veranstaltung trägt zu meinem Interesse am Thema bei.	○	○	○	○	○	☐
Der behandelte Stoff knüpft an meinen bisherigen Wissensstand an.	○	○	○	○	○	☐
Die Veranstaltung versetzt mich in die Lage, die Inhalte selbstständig zu vertiefen.	○	○	○	○	○	☐
Das fachliche Niveau der Veranstaltung empfinde ich als angemessen.	○	○	○	○	○	☐
Kommilitonen würde ich den Besuch dieser Veranstaltung empfehlen.	○	○	○	○	○	☐
Insgesamt gesehen, bin ich mit dieser Lehrveranstaltung zufrieden.	○	○	○	○	○	☐

Ich habe durch den Besuch dieser Lehrveranstaltung folgende Qualifikationen erworben:

	wenig			viel		k.A.
Wissen über Theorien und Modelle.	○	○	○	○	○	☐
Wissen über Fakten, Begriffe und Konzepte	○	○	○	○	○	☐
Wissen über Forschungsverfahren und wissenschaftliche Methoden.	○	○	○	○	○	☐
Anwendung von Theorien, Methoden, Konzepten.	○	○	○	○	○	☐
Praxiswissen, tätigkeitsrelevantes Wissen	○	○	○	○	○	☐
Schlüsselkompetenzen (Präsentieren, Arbeiten im Team, Recherchieren, ...) . . .	○	○	○	○	○	☐
Kompetenz zu unabhängigem und selbstständigem Arbeiten.	○	○	○	○	○	☐
Fachübergreifendes Denken	○	○	○	○	○	☐

	stimme nicht zu			stimme zu		k.A.
Insgesamt gesehen, bin ich mit den in dieser Veranstaltung erworbenen Qualifikationen zufrieden.	○	○	○	○	○	☐

Bitte beurteilen Sie die Rahmenbedingungen dieser Lehrveranstaltung.	stimme nicht zu			stimme zu		k.A.
Die räumlichen Gegebenheiten (Größe, bauliche Qualität, Lage, ...) sind für diese Veranstaltung ausreichend.	○	○	○	○	○	☐
Die Ausstattung (Medien, Technik, Modelle, ...) ist für diese Veranstaltung angemessen.	○	○	○	○	○	☐
Die Veranstaltung findet in einem angemessenen zeitlichen Rahmen (Zeitpunkt, Dauer, Überschneidungen, ...) statt.	○	○	○	○	○	☐
Begleitmaterialien (Literatur, Skript, ...) stehen in ausreichendem Maße zur Verfügung.	○	○	○	○	○	☐
Die verfügbaren Begleitmaterialien (Literatur, Skript, ...) sind hilfreich.	○	○	○	○	○	☐
Insgesamt gesehen, bin ich mit den Rahmenbedingungen dieser Lehrveranstaltung zufrieden.	○	○	○	○	○	☐

Der Dozent/die Dozentin...	stimme nicht zu				stimme zu	*k.A.*
hat Ziele und Struktur der Veranstaltung nachvollziehbar dargestellt.	○	○	○	○	○	☐
geht, soweit möglich, auf organisatorische Wünsche der Teilnehmenden ein.	○	○	○	○	○	☐
teilt die Veranstaltungszeit sinnvoll ein (auf Vortrag, Diskussion, Klärung von Fragen, ...).	○	○	○	○	○	☐
steht bei Bedarf für Rückfragen und weitere Hilfestellung zur Verfügung.	○	○	○	○	○	☐
schafft eine anregende Arbeitsatmosphäre.	○	○	○	○	○	☐
bereitet die Einzelsitzungen angemessen vor.	○	○	○	○	○	☐
greift inhaltliche Anregungen und Fragen der Teilnehmenden auf.	○	○	○	○	○	☐
ordnet Einzelaspekte in einen thematischen Gesamtzusammenhang ein.	○	○	○	○	○	☐
vermittelt auch komplizierte Inhalte klar und verständlich.	○	○	○	○	○	☐
setzt Präsentationsmedien und Visualisierung in hilfreicher Weise ein.	○	○	○	○	○	☐
gibt den Teilnehmenden in ausreichendem Maße Gelegenheit zur Diskussionsbeteiligung.	○	○	○	○	○	☐
strahlt Begeisterung für die vertretene Wissenschaft aus.	○	○	○	○	○	☐
Insgesamt gesehen, bin ich mit dem Beitrag des Dozenten/der Dozentin zu dieser Veranstaltung zufrieden.	○	○	○	○	○	☐

Die meisten Teilnehmenden dieser Lehrveranstaltung...	stimme nicht zu				stimme zu	*k.A.*
besuchen die Veranstaltung regelmäßig.	○	○	○	○	○	☐
bereiten sich auf die einzelnen Termine angemessen vor.	○	○	○	○	○	☐
beteiligen sich, soweit möglich, aktiv an der Veranstaltung.	○	○	○	○	○	☐
verfolgen die Veranstaltung aufmerksam und mit Interesse.	○	○	○	○	○	☐
Insgesamt gesehen, bin ich mit dem Verhalten der meisten Teilnehmenden zufrieden.	○	○	○	○	○	☐

Freie Items *(Festlegung erfolgt durch den Dozenten/die Dozentin)*						*k.A.*
Freies Item 1	○	○	○	○	○	☐
Freies Item 2	○	○	○	○	○	☐
Freies Item 3	○	○	○	○	○	☐

9541529285

Was hat Ihnen an dieser Veranstaltung besonders gut gefallen? *(Stichpunkte)*

Welche Anregungen oder Verbesserungsvorschläge haben Sie? *(Stichpunkte)*

Bitte beurteilen Sie abschließend diesen Fragebogen.	stimme nicht zu				stimme zu	*k.A.*
Dieser Fragebogen deckt die mir wichtigen Aspekte ausreichend ab.	○	○	○	○	○	☐
Die Fragen und Aussagen in diesem Fragebogen sind klar und verständlich formuliert.	○	○	○	○	○	☐

Abbildung A.2: Studentenfragebogen für Vorlesungen (Seite 2)

Friedrich-Schiller-Universität Jena
Universitätsprojekt Lehrevaluation

Lehrveranstaltungsevaluation

- Fragebogen für Vorträge und Vorlesungen (Dozent) -

Mit diesem Fragebogen nehmen Sie eine Einschätzung der von Ihnen durchgeführten Lehrveranstaltung **aus Ihrer Perspektive als Dozent oder Dozentin** vor. Im Ergebnisbericht werden Ihre Einschätzungen denen der Teilnehmenden gegenübergestellt und bieten so den ersten Ansatzpunkt für eine Diskussion.
Bitte kreuzen Sie jeweils die Antwortalternative an, die am ehesten Ihrer Wahrnehmung entspricht. Wenn eine Frage **nicht beantwortbar** ist oder auf die von Ihnen bewertete Veranstaltung **nicht anwendbar** ist, markieren Sie bitte '**keine Angabe**' (k.A.). Bitte senden Sie den ausgefüllten Dozentenbogen zusammen mit den von den Studierenden ausgefüllten Fragebögen zurück. Ihre Angaben werden zusammen mit den Studierendenfragebögen vom Universitätsprojekt **Lehrevaluation** ausgewertet und, wie alle Evaluationsdaten, vertraulich behandelt. Weitere Informationen zum Ablauf und zum Evaluationskonzept erhalten Sie im Internet unter **www.ule.uni-jena.de**.

Veranstaltung | **Dozent/-in** | **Datum**

Alter in Jahren ☐☐ Geschlecht ○ weiblich ○ männlich

Wie ist Ihr Vertragsverhältnis?
○ Daueranstellung mit Lehrpflicht ○ befristet aus Sondermitteln (Tutorien-Verträge o.ä.) ○ Anstellung ohne Lehrverpflichtung
○ Zeitvertrag mit Lehrpflicht (nach BAT o.ä.) ○ Lehrauftrag ○ anderes (auch ohne schriftlichen Vertrag)

Welcher ist Ihr höchster Abschluss?
○ Habilitation oder vergleichbar ○ Diplom/Magister/Staatsexamen o.ä. ○ anderes
○ Promotion ○ Zwischenprüfung abgeschlossen

Wie lange arbeiten Sie bereits in der Lehre?
○ erstmalig ○ 1 bis 2 Semester ○ 3 bis 4 Semester ○ länger als 4 Semester

Nicht in allen Veranstaltungen sollen gleichermaßen die verschiedenen Qualifikationen erworben werden, wie sie nachfolgend aufgeführt sind.
Legen Sie Ihre Planung zu Beginn des Semesters zugrunde und geben Sie an, welche Qualifikationen die Teilnehmenden im Laufe dieser Lehrveranstaltung (bisher) erwerben sollten.

	wenig				viel	k.A.
Wissen über Theorien und Modelle	○	○	○	○	○	☐
Wissen über Fakten, Begriffe und Konzepte	○	○	○	○	○	☐
Wissen über Forschungsverfahren und wissenschaftliche Methoden	○	○	○	○	○	☐
Anwendung von Theorien, Methoden, Konzepten	○	○	○	○	○	☐
Praxiswissen, tätigkeitsrelevantes Wissen	○	○	○	○	○	☐
Schlüsselkompetenzen (Präsentieren, Arbeiten im Team, Recherchieren, ...)	○	○	○	○	○	☐
Kompetenz zu unabhängigem und selbstständigem Arbeiten	○	○	○	○	○	☐
Fachübergreifendes Denken	○	○	○	○	○	☐

Neben dem Besuch der Veranstaltung zählen zur aktiven Teilnahme auch die Vor- und Nachbereitung der Sitzungen, Literaturarbeit und weitere Formen des Selbststudiums. **Wieviele Stunden pro Woche halten Sie an Selbststudium begleitend zu dieser Lehrveranstaltung für angemessen?** ☐☐,☐

Bitte treffen Sie eine Einschätzung zu diesen zusammenfassenden Aussagen über den tatsächlichen (bisherigen) Verlauf der Veranstaltung.	stimme nicht zu				stimme zu	k.A.
Die Veranstaltung trägt zum Interesse der Studierenden am Thema bei.	○	○	○	○	○	☐
Der behandelte Stoff knüpft an den bisherigen Wissensstand der Teilnehmenden an.	○	○	○	○	○	☐
Die Veranstaltung versetzt die Studierenden in die Lage, die Inhalte selbstständig zu vertiefen.	○	○	○	○	○	☐
Insgesamt gesehen, bin ich mit dieser Lehrveranstaltung zufrieden.	○	○	○	○	○	☐

2159431847

Abbildung A.3: Dozentenfragebogen für Vorlesungen (Seite 1)

Bitte schätzen Sie Ihr eigenes Verhalten als Dozent/Dozentin ein. Denken Sie dabei an typische Sitzungen im bisherigen Verlauf der Lehrveranstaltung. **In dieser Lehrveranstaltung...**

	wenig				viel	k.A.
habe ich Ziele und Struktur der Veranstaltung nachvollziehbar dargestellt.	○	○	○	○	○	☐
gehe ich, soweit möglich, auf organisatorische Wünsche der Teilnehmenden ein.	○	○	○	○	○	☐
teile ich die Veranstaltungszeit sinnvoll ein (auf Vortrag, Diskussion, Klärung von Fragen etc.).	○	○	○	○	○	☐
stehe ich bei Bedarf für Rückfragen und weitere Hilfestellung zur Verfügung.	○	○	○	○	○	☐
schaffe ich eine anregende Arbeitsatmosphäre.	○	○	○	○	○	☐
bereite ich die Einzelsitzung angemessen vor.	○	○	○	○	○	☐
greife ich inhaltliche Anregungen und Fragen der Teilnehmenden auf.	○	○	○	○	○	☐
ordne ich Einzelaspekte in einen thematischen Gesamtzusammenhang ein.	○	○	○	○	○	☐
vermittle ich auch komplizierte Inhalte klar und verständlich.	○	○	○	○	○	☐
setze ich Präsentationsmedien und Visualisierung in hilfreicher Weise ein.	○	○	○	○	○	☐
gebe ich den Teilnehmenden in ausreichendem Maße Gelegenheit zur Diskussionsbeteiligung.	○	○	○	○	○	☐
strahle ich Begeisterung für die eigene Wissenschaft aus.	○	○	○	○	○	☐

Die meisten Teilnehmenden dieser Lehrveranstaltung...	stimme nicht zu				stimme zu	k.A.
besuchen die Veranstaltung regelmäßig.	○	○	○	○	○	☐
bereiten sich auf die einzelnen Termine angemessen vor.	○	○	○	○	○	☐
beteiligen sich, soweit möglich, aktiv an der Veranstaltung.	○	○	○	○	○	☐
verfolgen die Veranstaltung aufmerksam und mit Interesse.	○	○	○	○	○	☐
Insgesamt gesehen, bin ich mit dem Verhalten der meisten Teilnehmenden zufrieden.	○	○	○	○	○	☐

Bitte beurteilen Sie die Rahmenbedingungen dieser Lehrveranstaltung.	stimme nicht zu				stimme zu	k.A.
Die räumlichen Gegebenheiten (Größe, bauliche Qualität, Lage, ...) sind für diese Veranstaltung ausreichend.	○	○	○	○	○	☐
Die Ausstattung (Medien, Technik, Modelle, ...) ist für diese Veranstaltung angemessen.	○	○	○	○	○	☐
Die Veranstaltung findet in einem angenehmen zeitlichen Rahmen (Zeitpunkt, Dauer, Überschneidungen) statt.	○	○	○	○	○	☐
Begleitmaterialien (Literatur, Skript, ...) stehen in ausreichendem Maße zur Verfügung.	○	○	○	○	○	☐
Die verfügbaren Begleitmaterialien (Literatur, Skript, ...) sind hilfreich.	○	○	○	○	○	☐
Insgesamt gesehen, bin ich mit den Rahmenbedingungen dieser Lehrveranstaltung zufrieden.	○	○	○	○	○	☐

Bitte beurteilen Sie abschließend diesen Fragebogen.	stimme nicht zu				stimme zu	k.A.
Dieser Fragebogen deckt die mir wichtigen Aspekte ausreichend ab.	○	○	○	○	○	☐
Die Fragen und Aussagen in diesem Fragebogen sind klar und verständlich formuliert.	○	○	○	○	○	☐

Falls Sie weitere Anmerkungen zu diesem Fragebogen oder zur Durchführung der Lehrveranstaltungsevaluation haben, können Sie diese hier notieren.

2046431842

Abbildung A.4: Dozentenfragebogen für Vorlesungen (Seite 2)

Abbildung A.5: Studentenfragebogen für Seminare (Seite 1)

Friedrich-Schiller-Universität Jena
seit 1558
Universitätsprojekt Lehrevaluation

Lehrveranstaltungsevaluation

- Fragebogen für Seminare und Veranstaltungen mit Teilnehmerbeiträgen (Teilnehmende) -

Mit diesem Fragebogen können Sie Ihren **persönlichen Eindruck** von dieser Lehrveranstaltung zurückmelden. Füllen Sie die zutreffende Antwortalternative bitte **mit einem dunklen Stift** (kein Bleistift) aus. Wenn eine Frage **nicht beantwortbar** ist oder Sie **keine Antwort geben möchten**, markieren Sie '**keine Angabe**' (k.A.). Möchten Sie eine **falsch markierte Antwort korrigieren,** streichen Sie diese bitte durch und markieren die von Ihnen gewünschte Antwortalternative. Die Fragebögen werden durch das Universitätsprojekt **Lehrevaluation (www.ule.uni-jena.de)** statistisch ausgewertet.

Veranstaltung **Dozent/-in** **Datum**

Bitte machen Sie für statistische Zwecke die folgenden Angaben.

Alter in Jahren ☐☐ Geschlecht ○ weiblich ○ männlich Fachsemester ☐☐

Bitte nennen Sie die Hauptgründe Ihres Veranstaltungsbesuches. (*Mehrfachnennungen möglich*)

○ inhaltliches Interesse ○ Pflichtveranstaltung ○ guter Ruf der Lehrkraft
○ keine Alternative verfügbar ○ zur Vorbereitung auf die Prüfung ○ andere Gründe

Ausschließlich für wissenschaftliche Fragestellungen bitten wir Sie, einen 5-stelligen Personencode anzugeben.

Der Personencode wird aus dem ersten Buchstaben Geburtsorts, dem zweiten Buchstaben Ihres Vornamens und dem dritten Buchstaben Ihres Nachnamens (ggf. Mädchenname) sowie den jeweils letzten Ziffern Ihres Geburtstags und Ihres Geburtsmonats gebildet. So lautet der Code für die aus Halle (1. Stelle) stammende Anja Krüger (2. und 3. Stelle), geb. am 26.03.1973 (4. und 5. Stelle des Codes): **HNÜ63**. ☐☐☐☐☐

An wievielen der bisherigen Termine dieser Veranstaltung haben Sie teilgenommen? . . . . . 0-20% ○ 21-40% ○ 41-60% ○ 61-80% ○ 81-100% ○

Wieviele **Stunden pro Woche** verbringen Sie durchschnittlich mit dem **Selbststudium** (bezogen auf alle Veranstaltungen **in diesem Semester**)? ☐☐,☐

Wieviele Stunden hiervon entfallen pro Woche auf **diese Veranstaltung**? ☐☐,☐

	stimme nicht zu			stimme zu		k.A.
Ich empfinde den von mir für diese Veranstaltung zu erbringenden Arbeitsaufwand als angemessen.	○	○	○	○	○	☐

5303018077

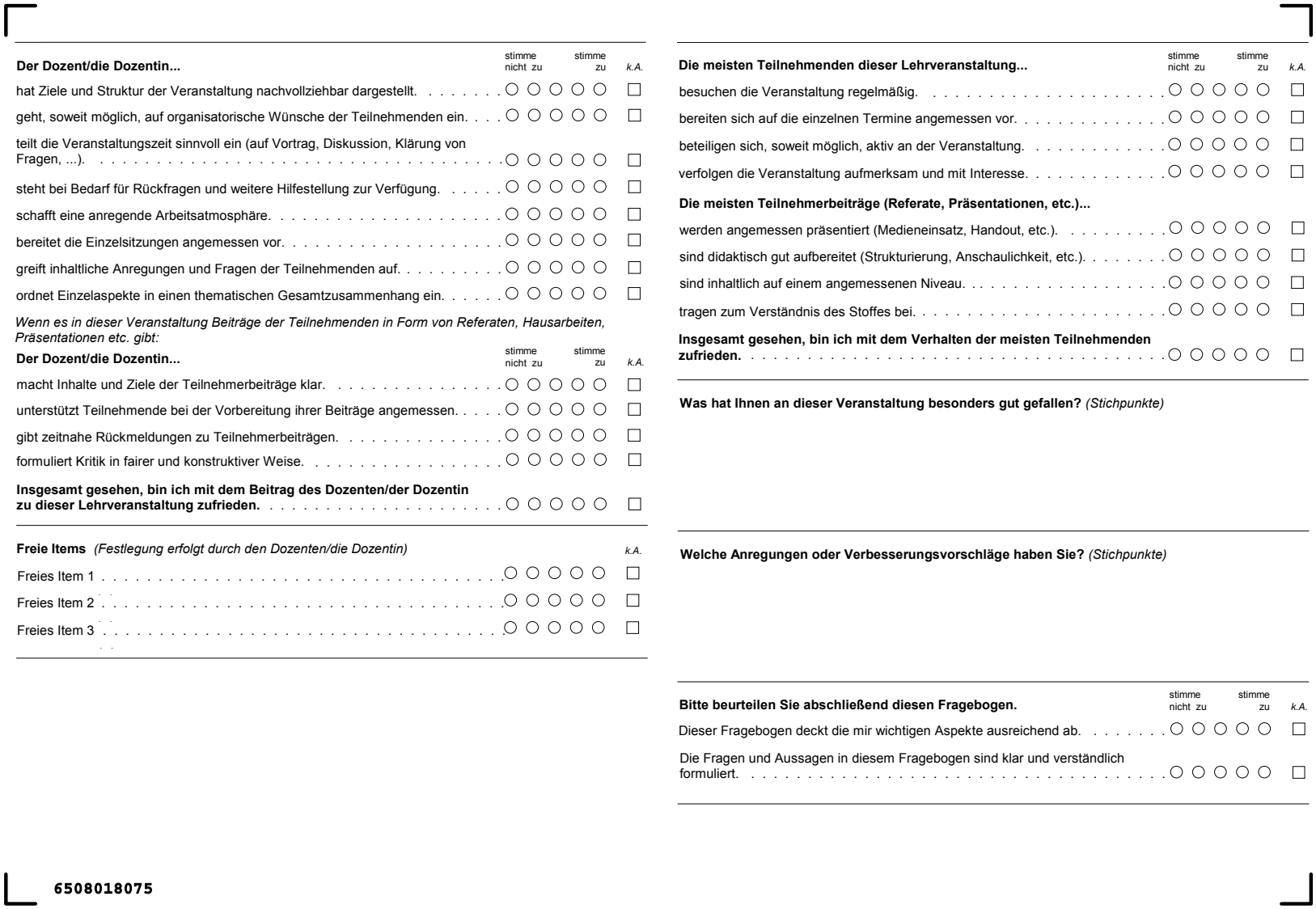
Bitte treffen Sie zunächst einige zusammenfassende Einschätzungen.	stimme nicht zu			stimme zu		k.A.
Die Veranstaltung trägt zu meinem Interesse am Thema bei.	○	○	○	○	○	☐
Der behandelte Stoff knüpft an meinen bisherigen Wissensstand an.	○	○	○	○	○	☐
Die Veranstaltung versetzt mich in die Lage, die Inhalte selbstständig zu vertiefen.	○	○	○	○	○	☐
Das fachliche Niveau der Veranstaltung empfinde ich als angemessen.	○	○	○	○	○	☐
Kommilitonen würde ich den Besuch dieser Veranstaltung empfehlen.	○	○	○	○	○	☐
Insgesamt gesehen, bin ich mit dieser Lehrveranstaltung zufrieden.	○	○	○	○	○	☐

Ich habe durch den Besuch dieser Lehrveranstaltung folgende Qualifikationen erworben:

	wenig			viel		k.A.
Wissen über Theorien und Modelle.	○	○	○	○	○	☐
Wissen über Fakten, Begriffe und Konzepte	○	○	○	○	○	☐
Wissen über Forschungsverfahren und wissenschaftliche Methoden.	○	○	○	○	○	☐
Anwendung von Theorien, Methoden, Konzepten.	○	○	○	○	○	☐
Praxiswissen, tätigkeitsrelevantes Wissen	○	○	○	○	○	☐
Schlüsselkompetenzen (Präsentieren, Arbeiten im Team, Recherchieren, ...) . . .	○	○	○	○	○	☐
Kompetenz zu unabhängigem und selbstständigem Arbeiten.	○	○	○	○	○	☐
Fachübergreifendes Denken	○	○	○	○	○	☐

	stimme nicht zu			stimme zu		k.A.
Insgesamt gesehen, bin ich mit den in dieser Veranstaltung erworbenen Qualifikationen zufrieden.	○	○	○	○	○	☐

Bitte beurteilen Sie die Rahmenbedingungen dieser Lehrveranstaltung.	stimme nicht zu			stimme zu		k.A.
Die räumlichen Gegebenheiten (Größe, bauliche Qualität, Lage, ...) sind für diese Veranstaltung ausreichend.	○	○	○	○	○	☐
Die Ausstattung (Medien, Technik, Modelle, ...) ist für diese Veranstaltung angemessen.	○	○	○	○	○	☐
Die Veranstaltung findet in einem angemessenen zeitlichen Rahmen (Zeitpunkt, Dauer, Überschneidungen, ...) statt.	○	○	○	○	○	☐
Begleitmaterialien (Literatur, Skript, ...) stehen in ausreichendem Maße zur Verfügung.	○	○	○	○	○	☐
Die verfügbaren Begleitmaterialien (Literatur, Skript, ...) sind hilfreich.	○	○	○	○	○	☐
Insgesamt gesehen, bin ich mit den Rahmenbedingungen dieser Lehrveranstaltung zufrieden.	○	○	○	○	○	☐



Der Dozent/die Dozentin...	stimme nicht zu				stimme zu	*k.A.*
hat Ziele und Struktur der Veranstaltung nachvollziehbar dargestellt.	○	○	○	○	○	☐
geht, soweit möglich, auf organisatorische Wünsche der Teilnehmenden ein.	○	○	○	○	○	☐
teilt die Veranstaltungszeit sinnvoll ein (auf Vortrag, Diskussion, Klärung von Fragen, ...).	○	○	○	○	○	☐
steht bei Bedarf für Rückfragen und weitere Hilfestellung zur Verfügung.	○	○	○	○	○	☐
schafft eine anregende Arbeitsatmosphäre.	○	○	○	○	○	☐
bereitet die Einzelsitzungen angemessen vor.	○	○	○	○	○	☐
greift inhaltliche Anregungen und Fragen der Teilnehmenden auf.	○	○	○	○	○	☐
ordnet Einzelaspekte in einen thematischen Gesamtzusammenhang ein.	○	○	○	○	○	☐

Wenn es in dieser Veranstaltung Beiträge der Teilnehmenden in Form von Referaten, Hausarbeiten, Präsentationen etc. gibt:

Der Dozent/die Dozentin...	stimme nicht zu				stimme zu	*k.A.*
macht Inhalte und Ziele der Teilnehmerbeiträge klar.	○	○	○	○	○	☐
unterstützt Teilnehmende bei der Vorbereitung ihrer Beiträge angemessen.	○	○	○	○	○	☐
gibt zeitnahe Rückmeldungen zu Teilnehmerbeiträgen.	○	○	○	○	○	☐
formuliert Kritik in fairer und konstruktiver Weise.	○	○	○	○	○	☐
Insgesamt gesehen, bin ich mit dem Beitrag des Dozenten/der Dozentin zu dieser Lehrveranstaltung zufrieden.	○	○	○	○	○	☐

Freie Items *(Festlegung erfolgt durch den Dozenten/die Dozentin)*						*k.A.*
Freies Item 1	○	○	○	○	○	☐
Freies Item 2	○	○	○	○	○	☐
Freies Item 3	○	○	○	○	○	☐

Die meisten Teilnehmenden dieser Lehrveranstaltung...	stimme nicht zu				stimme zu	*k.A.*
besuchen die Veranstaltung regelmäßig.	○	○	○	○	○	☐
bereiten sich auf die einzelnen Termine angemessen vor.	○	○	○	○	○	☐
beteiligen sich, soweit möglich, aktiv an der Veranstaltung.	○	○	○	○	○	☐
verfolgen die Veranstaltung aufmerksam und mit Interesse.	○	○	○	○	○	☐
Die meisten Teilnehmerbeiträge (Referate, Präsentationen, etc.)...						
werden angemessen präsentiert (Medieneinsatz, Handout, etc.).	○	○	○	○	○	☐
sind didaktisch gut aufbereitet (Strukturierung, Anschaulichkeit, etc.).	○	○	○	○	○	☐
sind inhaltlich auf einem angemessenen Niveau.	○	○	○	○	○	☐
tragen zum Verständnis des Stoffes bei.	○	○	○	○	○	☐
Insgesamt gesehen, bin ich mit dem Verhalten der meisten Teilnehmenden zufrieden.	○	○	○	○	○	☐

Was hat Ihnen an dieser Veranstaltung besonders gut gefallen? *(Stichpunkte)*

Welche Anregungen oder Verbesserungsvorschläge haben Sie? *(Stichpunkte)*

Bitte beurteilen Sie abschließend diesen Fragebogen.	stimme nicht zu				stimme zu	*k.A.*
Dieser Fragebogen deckt die mir wichtigen Aspekte ausreichend ab.	○	○	○	○	○	☐
Die Fragen und Aussagen in diesem Fragebogen sind klar und verständlich formuliert.	○	○	○	○	○	☐

6508018075

Abbildung A.6: Studentenfragebogen für Seminare (Seite 2)

Friedrich-Schiller-Universität Jena
Universitätsprojekt Lehrevaluation

Lehrveranstaltungsevaluation

- Fragebogen für Seminare und Veranstaltungen mit Teilnehmerbeiträgen (Dozent) -

Mit diesem Fragebogen nehmen Sie eine Einschätzung der von Ihnen durchgeführten Lehrveranstaltung **aus Ihrer Perspektive als Dozent oder Dozentin** vor. Im Ergebnisbericht werden Ihre Einschätzungen denen der Teilnehmenden gegenübergestellt und bieten so den ersten Ansatzpunkt für eine Diskussion.
Bitte kreuzen Sie jeweils die Antwortalternative an, die am ehesten Ihrer Wahrnehmung entspricht. Wenn eine Frage **nicht beantwortbar** ist oder auf die von Ihnen bewertete Veranstaltung **nicht anwendbar** ist, markieren Sie bitte '**keine Angabe**' (k.A.). Bitte senden Sie den ausgefüllten Dozentenbogen zusammen mit den von den Studierenden ausgefüllten Fragebögen zurück. Ihre Angaben werden zusammen mit den Studierendenfragebögen vom Universitätsprojekt **Lehrevaluation** ausgewertet und, wie alle Evaluationsdaten, vertraulich behandelt. Weitere Informationen zum Ablauf und zum Evaluationskonzept erhalten Sie im Internet unter **www.ule.uni-jena.de**.

Veranstaltung | **Dozent/-in** | **Datum**

Alter in Jahren ☐☐ Geschlecht ○ weiblich ○ männlich

Wie ist Ihr Vertragsverhältnis?
○ Daueranstellung mit Lehrpflicht ○ befristet aus Sondermitteln (Tutorien-Verträge o.ä.) ○ Anstellung ohne Lehrverpflichtung
○ Zeitvertrag mit Lehrpflicht (nach BAT o.ä.) ○ Lehrauftrag ○ anderes (auch ohne schriftlichen Vertrag)

Welcher ist Ihr höchster Abschluss?
○ Habilitation oder vergleichbar ○ Diplom/Magister/Staatsexamen o.ä. ○ anderes
○ Promotion ○ Zwischenprüfung abgeschlossen

Wie lange arbeiten Sie bereits in der Lehre?
○ erstmalig ○ 1 bis 2 Semester ○ 3 bis 4 Semester ○ länger als 4 Semester

Nicht in allen Veranstaltungen sollen gleichermaßen die verschiedenen Qualifikationen erworben werden, wie sie nachfolgend aufgeführt sind.
Legen Sie Ihre Planung zu Beginn des Semesters zugrunde und geben Sie an, welche Qualifikationen die Teilnehmenden im Laufe dieser Lehrveranstaltung (bisher) erworben sollten.

	wenig				viel	*k.A.*
Wissen über Theorien und Modelle	○	○	○	○	○	☐
Wissen über Fakten, Begriffe und Konzepte	○	○	○	○	○	☐
Wissen über Forschungsverfahren und wissenschaftliche Methoden	○	○	○	○	○	☐
Anwendung von Theorien, Methoden, Konzepten	○	○	○	○	○	☐
Praxiswissen, tätigkeitsrelevantes Wissen	○	○	○	○	○	☐
Schlüsselkompetenzen (Präsentieren, Arbeiten im Team, Recherchieren, ...)	○	○	○	○	○	☐
Kompetenz zu unabhängigem und selbstständigem Arbeiten	○	○	○	○	○	☐
Fachübergreifendes Denken	○	○	○	○	○	☐

Neben dem Besuch der Veranstaltung zählen zur aktiven Teilnahme auch die Vor- und Nachbereitung der Sitzungen, Literaturarbeit und weitere Formen des Selbststudiums. **Wieviele Stunden pro Woche halten Sie an Selbststudium begleitend zu dieser Lehrveranstaltung für angemessen?** ☐☐,☐

Bitte treffen Sie eine Einschätzung zu diesen zusammenfassenden Aussagen über den tatsächlichen (bisherigen) Verlauf der Veranstaltung.

	stimme nicht zu				stimme zu	*k.A.*
Die Veranstaltung trägt zum Interesse der Studierenden am Thema bei.	○	○	○	○	○	☐
Der behandelte Stoff knüpft an den bisherigen Wissensstand der Teilnehmenden an.	○	○	○	○	○	☐
Die Veranstaltung versetzt die Studierenden in die Lage, die Inhalte selbstständig zu vertiefen.	○	○	○	○	○	☐
Insgesamt gesehen, bin ich mit dieser Lehrveranstaltung zufrieden.	○	○	○	○	○	☐

7304534196

Abbildung A.7: Dozentenfragebogen für Seminare (Seite 1)

Bitte schätzen Sie Ihr eigenes Verhalten als Dozent/Dozentin ein. Denken Sie dabei an typische Sitzungen im bisherigen Verlauf der Lehrveranstaltung. **In dieser Lehrveranstaltung...**

	wenig				viel	k.A.
habe ich Ziele und Struktur der Veranstaltung nachvollziehbar dargestellt.	○	○	○	○	○	☐
gehe ich, soweit möglich, auf organisatorische Wünsche der Teilnehmenden ein.	○	○	○	○	○	☐
teile ich die Veranstaltungszeit sinnvoll ein (auf Vortrag, Diskussion, Klärung von Fragen etc.).	○	○	○	○	○	☐
stehe ich bei Bedarf für Rückfragen und weitere Hilfestellung zur Verfügung.	○	○	○	○	○	☐
schaffe ich eine anregende Arbeitsatmosphäre.	○	○	○	○	○	☐
bereite ich die Einzelsitzung angemessen vor.	○	○	○	○	○	☐
greife ich inhaltliche Anregungen und Fragen der Teilnehmenden auf.	○	○	○	○	○	☐
ordne ich Einzelaspekte in einen thematischen Gesamtzusammenhang ein.	○	○	○	○	○	☐

Wenn es in dieser Veranstaltung Beiträge der Teilnehmenden in Form von Referaten, Hausarbeiten, Präsentationen etc. gibt:

In dieser Veranstaltung...	stimme nicht zu				stimme zu	k.A.
mache ich Inhalte und Ziele der Teilnehmerbeiträge klar.	○	○	○	○	○	☐
unterstütze ich Teilnehmende bei der Vorbereitung ihrer Beiträge angemessen.	○	○	○	○	○	☐
gebe ich zeitnahe Rückmeldung zu Teilnehmerbeiträgen.	○	○	○	○	○	☐
formuliere ich Kritik in fairer und konstruktiver Weise.	○	○	○	○	○	☐

Die meisten Teilnehmenden dieser Lehrveranstaltung...	stimme nicht zu				stimme zu	k.A.
besuchen die Veranstaltung regelmäßig.	○	○	○	○	○	☐
bereiten sich auf die einzelnen Termine angemessen vor.	○	○	○	○	○	☐
beteiligen sich, soweit möglich, aktiv an der Veranstaltung.	○	○	○	○	○	☐
verfolgen die Veranstaltung aufmerksam und mit Interesse.	○	○	○	○	○	☐
Die meisten Teilnehmerbeiträge (Referate, Präsentationen, etc.)...						
werden angemessen präsentiert (Medieneinsatz, Handout, etc.).	○	○	○	○	○	☐
sind didaktisch gut aufbereitet (Strukturierung, Anschaulichkeit, etc.).	○	○	○	○	○	☐
sind inhaltlich auf einem angemessenen Niveau.	○	○	○	○	○	☐
tragen zum Verständnis des Stoffes bei.	○	○	○	○	○	☐
Insgesamt gesehen, bin ich mit dem Verhalten der meisten Teilnehmenden zufrieden.	○	○	○	○	○	☐

Bitte beurteilen Sie die Rahmenbedingungen dieser Lehrveranstaltung.	stimme nicht zu				stimme zu	k.A.
Die räumlichen Gegebenheiten (Größe, bauliche Qualität, Lage, ...) sind für diese Veranstaltung ausreichend.	○	○	○	○	○	☐
Die Ausstattung (Medien, Technik, Modelle, ...) ist für diese Veranstaltung angemessen.	○	○	○	○	○	☐
Die Veranstaltung findet in einem angenehmen zeitlichen Rahmen (Zeitpunkt, Dauer, Überschneidungen) statt.	○	○	○	○	○	☐
Begleitmaterialien (Literatur, Skript, ...) stehen in ausreichendem Maße zur Verfügung.	○	○	○	○	○	☐
Die verfügbaren Begleitmaterialien (Literatur, Skript, ...) sind hilfreich.	○	○	○	○	○	☐
Insgesamt gesehen, bin ich mit den Rahmenbedingungen dieser Lehrveranstaltung zufrieden.	○	○	○	○	○	☐

Bitte beurteilen Sie abschließend diesen Fragebogen.	stimme nicht zu				stimme zu	k.A.
Dieser Fragebogen deckt die mir wichtigen Aspekte ausreichend ab.	○	○	○	○	○	☐
Die Fragen und Aussagen in diesem Fragebogen sind klar und verständlich formuliert.	○	○	○	○	○	☐

Falls Sie weitere Anmerkungen zu diesem Fragebogen oder zur Durchführung der Lehrveranstaltungsevaluation haben, können Sie diese hier notieren.

6769534190

Abbildung A.8: Dozentenfragebogen für Seminare (Seite 2)

Abbildung A.9: Studentenfragebogen für Übungen(Seite 1)

seit 1558

Friedrich-Schiller-Universität Jena

Universitätsprojekt Lehrevaluation

Lehrveranstaltungsevaluation

- Fragebogen für Übungen und Praxisveranstaltungen (Teilnehmende) -

Mit diesem Fragebogen können Sie Ihren **persönlichen Eindruck** von dieser Lehrveranstaltung zurückmelden. Füllen Sie die zutreffende Antwortalternative bitte **mit einem dunklen Stift** (kein Bleistift) aus. Wenn eine Frage **nicht beantwortbar** ist oder Sie **keine Antwort geben möchten**, markieren Sie '**keine Angabe**' (k.A.). Möchten Sie eine **falsch markierte Antwort korrigieren,** streichen Sie diese bitte durch und markieren die von Ihnen gewünschte Antwortalternative. Die Fragebögen werden durch das Universitätsprojekt **Lehrevaluation (www.ule.uni-jena.de)** statistisch ausgewertet.

Veranstaltung **Dozent/-in** **Datum**

Bitte machen Sie für statistische Zwecke die folgenden Angaben.

Alter in Jahren ☐☐ Geschlecht ○ weiblich ○ männlich Fachsemester ☐☐

Bitte nennen Sie die Hauptgründe Ihres Veranstaltungsbesuches. (*Mehrfachnennungen möglich*)

○ inhaltliches Interesse ○ Pflichtveranstaltung ○ guter Ruf der Lehrkraft
○ keine Alternative verfügbar ○ zur Vorbereitung auf die Prüfung ○ andere Gründe

Ausschließlich für wissenschaftliche Fragestellungen bitten wir Sie, einen 5-stelligen Personencode anzugeben.

Der Personencode wird aus dem ersten Buchstaben Geburtsorts, dem zweiten Buchstaben Ihres Vornamens und dem dritten Buchstaben Ihres Nachnamens (ggf. Mädchenname) sowie den jeweils letzten Ziffern Ihres Geburtstags und Ihres Geburtsmonats gebildet. So lautet der Code für die aus Halle (1. Stelle) stammende Anja Krüger (2. und 3. Stelle), geb. am 26.03.1973 (4. und 5. Stelle des Codes): **HNÜ63**. ☐☐☐☐☐

An wievielen der bisherigen Termine dieser Veranstaltung haben Sie teilgenommen? . . . . . 0-20% ○ 21-40% ○ 41-60% ○ 61-80% ○ 81-100% ○

Wieviele **Stunden pro Woche** verbringen Sie durchschnittlich mit dem **Selbststudium** (bezogen auf alle Veranstaltungen **in diesem Semester**)? ☐☐,☐

Wieviele Stunden hiervon entfallen pro Woche auf **diese Veranstaltung**? ☐☐,☐

Ich empfinde den von mir für diese Veranstaltung zu erbringenden Arbeitsaufwand als angemessen.	stimme nicht zu ○ ○ ○ ○ ○ stimme zu	k.A. ☐

9141485068

Bitte treffen Sie zunächst einige zusammenfassende Einschätzungen.	stimme nicht zu ○ ○ ○ ○ ○ stimme zu	k.A.
Die Veranstaltung trägt zu meinem Interesse am Thema bei.		☐
Der behandelte Stoff knüpft an meinen bisherigen Wissensstand an.		☐
Die Veranstaltung versetzt mich in die Lage, die Inhalte selbstständig zu vertiefen.		☐
Das fachliche Niveau der Veranstaltung empfinde ich als angemessen.		☐
Kommilitonen würde ich den Besuch dieser Veranstaltung empfehlen.		☐
Insgesamt gesehen, bin ich mit dieser Lehrveranstaltung zufrieden.		☐

Ich habe durch den Besuch dieser Lehrveranstaltung folgende Qualifikationen erworben:

	wenig ○ ○ ○ ○ ○ viel	k.A.
Wissen über Theorien und Modelle		☐
Wissen über Fakten, Begriffe und Konzepte		☐
Wissen über Forschungsverfahren und wissenschaftliche Methoden		☐
Anwendung von Theorien, Methoden, Konzepten		☐
Praxiswissen, tätigkeitsrelevantes Wissen		☐
Schlüsselkompetenzen (Präsentieren, Arbeiten im Team, Recherchieren, ...)		☐
Kompetenz zu unabhängigem und selbstständigem Arbeiten.		☐
Fachübergreifendes Denken		☐
Insgesamt gesehen, bin ich mit den in dieser Veranstaltung erworbenen Qualifikationen zufrieden.	stimme nicht zu ○ ○ ○ ○ ○ stimme zu	k.A. ☐

Bitte beurteilen Sie die Rahmenbedingungen dieser Lehrveranstaltung.	stimme nicht zu ○ ○ ○ ○ ○ stimme zu	k.A.
Die räumlichen Gegebenheiten (Größe, bauliche Qualität, Lage, ...) sind für diese Veranstaltung ausreichend.		☐
Die Ausstattung (Medien, Technik, Modelle, ...) ist für diese Veranstaltung angemessen.		☐
Die Veranstaltung findet in einem angemessenen zeitlichen Rahmen (Zeitpunkt, Dauer, Überschneidungen, ...) statt.		☐
Begleitmaterialien (Literatur, Skript, ...) stehen in ausreichendem Maße zur Verfügung.		☐
Die verfügbaren Begleitmaterialien (Literatur, Skript, ...) sind hilfreich.		☐
Insgesamt gesehen, bin ich mit den Rahmenbedingungen dieser Lehrveranstaltung zufrieden.		☐

Abbildung A.10: Studentenfragebogen für Übungen (Seite 2)

Der Dozent/die Dozentin...	stimme nicht zu				stimme zu	*k.A.*
hat Ziele und Struktur der Veranstaltung nachvollziehbar dargestellt.	○	○	○	○	○	☐
geht, soweit möglich, auf organisatorische Wünsche der Teilnehmenden ein.	○	○	○	○	○	☐
teilt die Veranstaltungszeit sinnvoll ein (auf Vortrag, Diskussion, Klärung von Fragen, ...).	○	○	○	○	○	☐
steht bei Bedarf für Rückfragen und weitere Hilfestellung zur Verfügung.	○	○	○	○	○	☐
schafft eine anregende Arbeitsatmosphäre.	○	○	○	○	○	☐
bereitet die Einzelsitzungen angemessen vor.	○	○	○	○	○	☐
greift inhaltliche Anregungen und Fragen der Teilnehmenden auf.	○	○	○	○	○	☐
ordnet Einzelaspekte in einen thematischen Gesamtzusammenhang ein.	○	○	○	○	○	☐

Wenn im Rahmen dieser Veranstaltung durch die Teilnehmenden Aufgaben, praktische Übungen etc. durchzuführen oder zu bearbeiten sind:

Der Dozent/die Dozentin...	stimme nicht zu				stimme zu	*k.A.*
unterstützt die Teilnehmenden angemessen bei der Bearbeitung der Aufgaben.	○	○	○	○	○	☐
wertet die Ergebnisse und Lösungen ausführlich mit den Teilnehmenden aus.	○	○	○	○	○	☐
geht auf die Ergebnisse so ein, dass aus Fehlern gelernt werden kann.	○	○	○	○	○	☐
Insgesamt gesehen, bin ich mit dem Beitrag des Dozenten/der Dozentin zu dieser Veranstaltung zufrieden.	○	○	○	○	○	☐

Wenn im Rahmen dieser Veranstaltung durch die Teilnehmenden Aufgaben, praktische Übungen etc. durchzuführen oder zu bearbeiten sind:

Die Aufgaben, praktischen Übungen etc. ...	stimme nicht zu				stimme zu	*k.A.*
sind verständlich formuliert.	○	○	○	○	○	☐
sind von angemessenem Schwierigkeitsgrad.	○	○	○	○	○	☐
stehen in angemessener Anzahl zur Verfügung.	○	○	○	○	○	☐
ermöglichen den Teilnehmenden, die für sie wichtigen Aspekte zu vertiefen.	○	○	○	○	○	☐

Freie Items *(Festlegung erfolgt durch den Dozenten/die Dozentin)*						*k.A.*
Freies Item 1	○	○	○	○	○	☐
Freies Item 2	○	○	○	○	○	☐
Freies Item 3	○	○	○	○	○	☐

5471485060

Die meisten Teilnehmenden dieser Lehrveranstaltung...	stimme nicht zu				stimme zu	*k.A.*
besuchen die Veranstaltung regelmäßig.	○	○	○	○	○	☐
bereiten sich auf die einzelnen Termine angemessen vor.	○	○	○	○	○	☐
beteiligen sich, soweit möglich, aktiv an der Veranstaltung.	○	○	○	○	○	☐
verfolgen die Veranstaltung aufmerksam und mit Interesse.	○	○	○	○	○	☐
Insgesamt gesehen, bin ich mit dem Verhalten der meisten Teilnehmenden zufrieden.	○	○	○	○	○	☐

Was hat Ihnen an dieser Veranstaltung besonders gut gefallen? *(Stichpunkte)*

Welche Anregungen oder Verbesserungsvorschläge haben Sie? *(Stichpunkte)*

Bitte beurteilen Sie abschließend diesen Fragebogen.	stimme nicht zu				stimme zu	*k.A.*
Dieser Fragebogen deckt die mir wichtigen Aspekte ausreichend ab.	○	○	○	○	○	☐
Die Fragen und Aussagen in diesem Fragebogen sind klar und verständlich formuliert.	○	○	○	○	○	☐

Friedrich-Schiller-Universität Jena
Universitätsprojekt Lehrevaluation

Lehrveranstaltungsevaluation

- Fragebogen für Übungen und Praxisveranstaltungen (Dozent) -

Mit diesem Fragebogen nehmen Sie eine Einschätzung der von Ihnen durchgeführten Lehrveranstaltung **aus Ihrer Perspektive als Dozent oder Dozentin** vor. Im Ergebnisbericht werden Ihre Einschätzungen denen der Teilnehmenden gegenübergestellt und bieten so den ersten Ansatzpunkt für eine Diskussion.
Bitte kreuzen Sie jeweils die Antwortalternative an, die am ehesten Ihrer Wahrnehmung entspricht. Wenn eine Frage **nicht beantwortbar** ist oder auf die von Ihnen bewertete Veranstaltung **nicht anwendbar** ist, markieren Sie bitte '**keine Angabe**' (k.A.). Bitte senden Sie den ausgefüllten Dozentenbogen zusammen mit den von den Studierenden ausgefüllten Fragebögen zurück. Ihre Angaben werden zusammen mit den Studierendenfragebögen vom **Universitätsprojekt Lehrevaluation** ausgewertet und, wie alle Evaluationsdaten, vertraulich behandelt. Weitere Informationen zum Ablauf und zum Evaluationskonzept erhalten Sie im Internet unter **www.ule.uni-jena.de**.

Veranstaltung | **Dozent/-in** | **Datum**

Alter in Jahren ☐☐ Geschlecht ○ weiblich ○ männlich

Wie ist Ihr Vertragsverhältnis?
○ Daueranstellung mit Lehrpflicht ○ befristet aus Sondermitteln (Tutorien-Verträge o.ä.) ○ Anstellung ohne Lehrverpflichtung
○ Zeitvertrag mit Lehrpflicht (nach BAT o.ä.) ○ Lehrauftrag ○ anderes (auch ohne schriftlichen Vertrag)

Welcher ist Ihr höchster Abschluss?
○ Habilitation oder vergleichbar ○ Diplom/Magister/Staatsexamen o.ä. ○ anderes
○ Promotion ○ Zwischenprüfung abgeschlossen

Wie lange arbeiten Sie bereits in der Lehre?
○ erstmalig ○ 1 bis 2 Semester ○ 3 bis 4 Semester ○ länger als 4 Semester

Nicht in allen Veranstaltungen sollen gleichermaßen die verschiedenen Qualifikationen erworben werden, wie sie nachfolgend aufgeführt sind.
Legen Sie Ihre Planung zu Beginn des Semesters zugrunde und geben Sie an, welche Qualifikationen die Teilnehmenden im Laufe dieser Lehrveranstaltung (bisher) erwerben sollten.

	wenig				viel	k.A.
Wissen über Theorien und Modelle	○	○	○	○	○	☐
Wissen über Fakten, Begriffe und Konzepte	○	○	○	○	○	☐
Wissen über Forschungsverfahren und wissenschaftliche Methoden	○	○	○	○	○	☐
Anwendung von Theorien, Methoden, Konzepten	○	○	○	○	○	☐
Praxiswissen, tätigkeitsrelevantes Wissen	○	○	○	○	○	☐
Schlüsselkompetenzen (Präsentieren, Arbeiten im Team, Recherchieren, ...)	○	○	○	○	○	☐
Kompetenz zu unabhängigem und selbstständigem Arbeiten	○	○	○	○	○	☐
Fachübergreifendes Denken	○	○	○	○	○	☐

Neben dem Besuch der Veranstaltung zählen zur aktiven Teilnahme auch die Vor- und Nachbereitung der Sitzungen, Literaturarbeit und weitere Formen des Selbststudiums. **Wieviele Stunden pro Woche halten Sie an Selbststudium begleitend zu dieser Lehrveranstaltung für angemessen?** ☐☐,☐

Bitte treffen Sie eine Einschätzung zu diesen zusammenfassenden Aussagen über den tatsächlichen (bisherigen) Verlauf der Veranstaltung.	stimme nicht zu				stimme zu	k.A.
Die Veranstaltung trägt zum Interesse der Studierenden am Thema bei.	○	○	○	○	○	☐
Der behandelte Stoff knüpft an den bisherigen Wissensstand der Teilnehmenden an.	○	○	○	○	○	☐
Die Veranstaltung versetzt die Studierenden in die Lage, die Inhalte selbstständig zu vertiefen.	○	○	○	○	○	☐
Insgesamt gesehen, bin ich mit dieser Lehrveranstaltung zufrieden.	○	○	○	○	○	☐

4768468968

Abbildung A.11: Dozentenfragebogen für Übungen (Seite 1)

Bitte schätzen Sie Ihr eigenes Verhalten als Dozent/Dozentin ein. Denken Sie dabei an typische Sitzungen im bisherigen Verlauf der Lehrveranstaltung. **In dieser Lehrveranstaltung...**

	wenig				viel	k.A.
habe ich Ziele und Struktur der Veranstaltung nachvollziehbar dargestellt.	○	○	○	○	○	☐
gehe ich, soweit möglich, auf organisatorische Wünsche der Teilnehmenden ein.	○	○	○	○	○	☐
teile ich die Veranstaltungszeit sinnvoll ein (auf Vortrag, Diskussion, Klärung von Fragen etc.).	○	○	○	○	○	☐
stehe ich bei Bedarf für Rückfragen und weitere Hilfestellung zur Verfügung.	○	○	○	○	○	☐
schaffe ich eine anregende Arbeitsatmosphäre.	○	○	○	○	○	☐
bereite ich die Einzelsitzung angemessen vor.	○	○	○	○	○	☐
greife ich inhaltliche Anregungen und Fragen der Teilnehmenden auf.	○	○	○	○	○	☐
ordne ich Einzelaspekte in einen thematischen Gesamtzusammenhang ein.	○	○	○	○	○	☐

Wenn im Rahmen dieser Veranstaltung durch die Teilnehmenden Aufgaben, praktische Übungen etc. durchzuführen oder zu bearbeiten sind:

In dieser Veranstaltung...	stimme nicht zu				stimme zu	k.A.
unterstütze ich Teilnehmende angemessen bei der Bearbeitung der Aufgaben.	○	○	○	○	○	☐
werte ich die Ergebnisse und Lösungen ausführlich mit den Teilnehmenden aus.	○	○	○	○	○	☐
gehe ich auf die Ergebnisse so ein, dass aus Fehlern gelernt werden kann.	○	○	○	○	○	☐
Die Aufgaben, praktischen Übungen etc. ...						
sind verständlich formuliert.	○	○	○	○	○	☐
sind von angemessenem Schwierigkeitsgrad.	○	○	○	○	○	☐
stehen in angemessener Anzahl zur Verfügung.	○	○	○	○	○	☐
ermöglichen den Teilnehmenden, die für sie wichtigen Aspekte zu vertiefen.	○	○	○	○	○	☐

Die meisten Teilnehmenden dieser Lehrveranstaltung...	stimme nicht zu				stimme zu	k.A.
besuchen die Veranstaltung regelmäßig.	○	○	○	○	○	☐
bereiten sich auf die einzelnen Termine angemessen vor.	○	○	○	○	○	☐
beteiligen sich, soweit möglich, aktiv an der Veranstaltung.	○	○	○	○	○	☐
verfolgen die Veranstaltung aufmerksam und mit Interesse.	○	○	○	○	○	☐
Insgesamt gesehen, bin ich mit dem Verhalten der meisten Teilnehmenden zufrieden.	○	○	○	○	○	☐

Bitte beurteilen Sie die Rahmenbedingungen dieser Lehrveranstaltung.	stimme nicht zu				stimme zu	k.A.
Die räumlichen Gegebenheiten (Größe, bauliche Qualität, Lage, ...) sind für diese Veranstaltung ausreichend.	○	○	○	○	○	☐
Die Ausstattung (Medien, Technik, Modelle, ...) ist für diese Veranstaltung angemessen.	○	○	○	○	○	☐
Die Veranstaltung findet in einem angenehmen zeitlichen Rahmen (Zeitpunkt, Dauer, Überschneidungen) statt.	○	○	○	○	○	☐
Begleitmaterialien (Literatur, Skript, ...) stehen in ausreichendem Maße zur Verfügung.	○	○	○	○	○	☐
Die verfügbaren Begleitmaterialien (Literatur, Skript, ...) sind hilfreich.	○	○	○	○	○	☐
Insgesamt gesehen, bin ich mit den Rahmenbedingungen dieser Lehrveranstaltung zufrieden.	○	○	○	○	○	☐

Bitte beurteilen Sie abschließend diesen Fragebogen.	stimme nicht zu				stimme zu	k.A.
Dieser Fragebogen deckt die mir wichtigen Aspekte ausreichend ab.	○	○	○	○	○	☐
Die Fragen und Aussagen in diesem Fragebogen sind klar und verständlich formuliert.	○	○	○	○	○	☐

Falls Sie weitere Anmerkungen zu diesem Fragebogen oder zur Durchführung der Lehrveranstaltungsevaluation haben, können Sie diese hier notieren.

7220468964

Abbildung A.12: Dozentenfragebogen für Übungen (Seite 2)

Ehrenwörtliche Erklärung

Hiermit erkläre ich, dass mir die geltende Promotionsordnung der Fakultät für Sozial- und Verhaltenswissenschaften der Friedrich-Schiller-Universität Jena vom 06.05.2009 mit den Änderungen vom 17.11.2010, 19.06.2012 und 21.01.2014 bekannt ist. Ich habe die Dissertation selbst angefertigt, keine Textabschnitte eines Dritten oder eigener Prüfungsarbeiten ohne Kennzeichnung übernommen und alle von mir benutzten Hilfsmittel, persönlichen Mitteilungen und Quellen in meiner Arbeit angegeben. Bei der Überarbeitung der Artikel und des Manuskriptes wurde ich vornehmlich von Marie-Ann Sengewald, vom Koautor der Artikel Anja Vetterlein und von Prof. Dr. Steyer unentgeltlich unterstützt. Alexander Schauerte, Elisa Wiedmann und Simon Schmitt haben im Rahmen ihrer Tätigkeit als studentische Hilfskraft am Lehrstuhl für Methodenlehre und Evaluationsforschung die Arbeit in Teilen auf orthografische Korrektheit geprüft. Ich habe keine Hilfe eines Promotionsberaters in Anspruch genommen und Dritte haben weder unmittelbar oder mittelbar geldwerte Leistungen von mir für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen. Die Dissertation habe ich nicht als Promotionsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht. Darüber hinaus habe ich keine gleiche oder in wesentlichen Teilen ähnliche Dissertation oder eine andere Abhandlung bei einer anderen Hochschule oder anderen Fakultät als Dissertation eingereicht.

Fürth, den 04. Februar 2016